

Galaxy-based multi-stage two-step database searching pipeline for improved multi-omics analysis

Praveen Kumar¹, James E. Johnson², Thomas McGowan², Matthew C. Chambers⁴, Mohammad Heydarian⁵, Subina Mehta³, Caleb Easterly³, Shane Hubler⁶, Joel D. Rudney⁷, Pratik Jagtap³, Timothy J. Griffin³

¹Bioinformatics and Computational Biology, University of Minnesota, Minneapolis, Minnesota

²Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota

³Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, Minnesota

⁴Department of Biochemistry, Vanderbilt University, Nashville, Tennessee

⁵Department of Biology, Johns Hopkins University, Baltimore, Maryland

⁶DRhapsody Data, LLC, Madison, Wisconsin

⁷Department of Diagnostic and Biological Sciences, School of Dentistry, University of Minnesota, Minneapolis, Minnesota



Introduction

- In a proteogenomic study, genome/transcriptome sequencing data is used for the generation of potentially expressed protein variants, producing a large number of possible sequences to be matched with (MS/MS) data.
- Matching the MS/MS data to peptide sequences contained within a database indicates the expression of variant proteins in the sample.
- One of the challenges that proteogenomics analysis faces is the large databases, where translated RNA-Seq database are usually appended to the database of known proteins that have been previously characterized (Figure 1).
- Increasing database size increases false-positive identifications and loss of sensitivity for identifying true peptide spectrum matches (PSMs) (Figure 2).
- A metaproteomics study encounters similar challenge of large databases as the database being used contains protein sequences from thousands of microbes.
- We developed the multi-stage two-step database searching method (Figure 3), which first uses a multi-stage searching routine to reduce the size of the database, and then match the MS/MS to the reduced database in the second step.
- We also proposed and tested sectioning method for metaproteomic study, where we section the large microbial proteome database and used the multi-stage two-step database searching method, which helped in reducing the database size and observed increase in the number of PSM identifications.
- This implementation is available as a workflow on Galaxy-P that can be shared with other researchers and integrated in other pipelines. These workflows are available at:
 - <https://galaxy.msi.umn.edu> (hosted by Minnesota Supercomputing Institute)
 - <http://galaxy-proteogenomics.duckdns.org> (hosted by JetStream)

Proteogenomics Data

- Samples from mouse early developmental B cells (RNA-Seq and MS/MS) – Heydarian et al. (2014)
- MS/MS data searched against following databases (Table 1):
 - UniProt mouse protein sequences
 - 3-frame translated cDNA (from Ensembl)
 - Splice junction variants (derived from RNA-Seq data)
 - 3-frame translated long non-coding RNA (derived from RNA-Seq data)
 - Single amino-acid variants (derived from RNA-Seq data)

Table 1

Database	Number of entries in the database	
	All DB combined (Traditional Method)	Multi-stage two-step
UniProt mouse + Contaminants	59,605	15,987
3-frame translated cDNA	2,110,551	19,358
Splice junction variants	64,917	10,964
Long non-coding RNA	420,890	17,373
Single amino acid variations	22182	9,999
Total	2,673,777	73,681

Table 1: Number of protein sequences included in each database. Reduced database in multi-stage two-step method was 35 times smaller than the traditional database.

Metaproteomics Data

- Proteins from dental caries samples grown *in vitro* – Rudney, J.D. et al. (2015)
- Database: Human Oral Microbiome Database (HOMD) - 1,079,626 sequences
- Traditional method:
 - Database: All HOMD sequences
 - Sectioning method:
 - 5 sections of HOMD (approx. 200,000 sequences each)
 - Multi-stage two-step method
 - Reduced database size (Two-step): 129,562

Figure 7

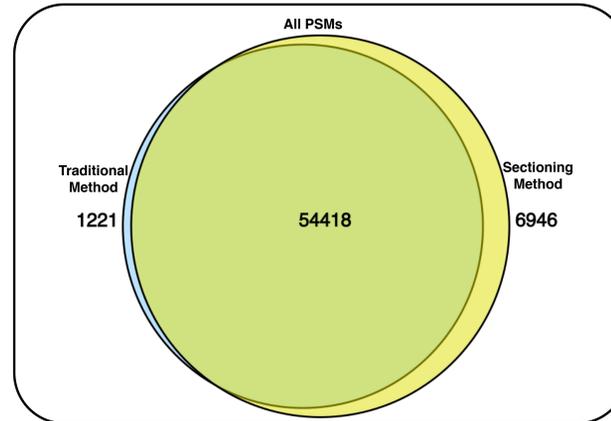


Figure 7: Comparing the number of all qualified PSMs identified by each method.

Figure 8

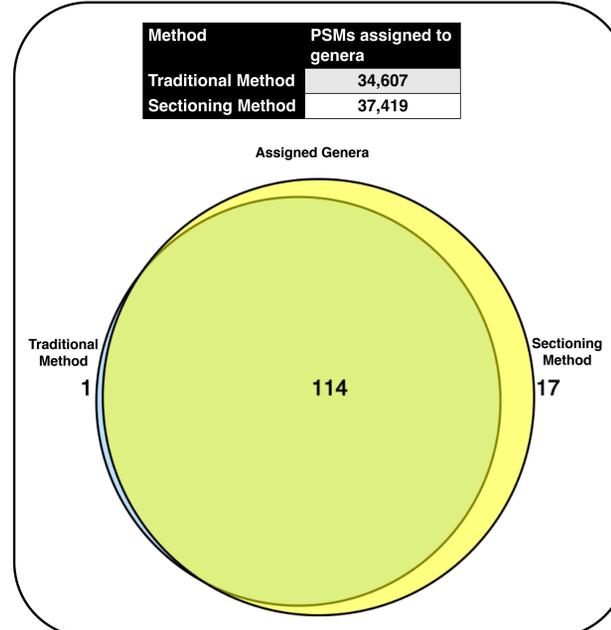


Figure 8: Comparing the number of genera assigned by the PSMs identified by each method (Table shows number of PSMs that were assigned to genera)

Figure 9

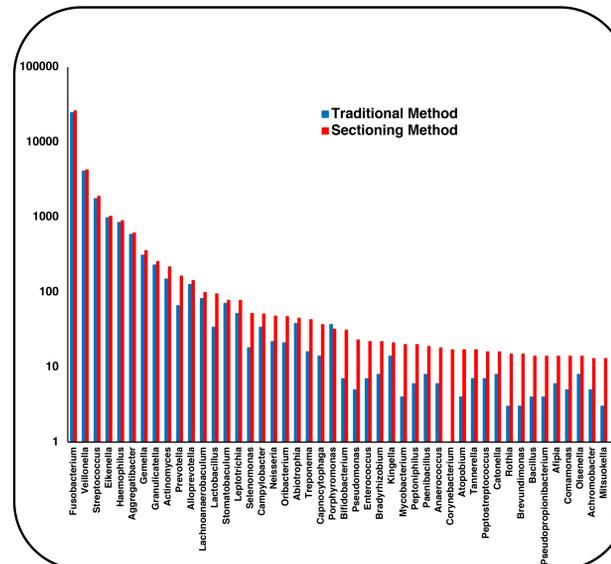


Figure 9: Number of PSMs from each method that were assigned to each genera from Unipect analysis

Figure 10

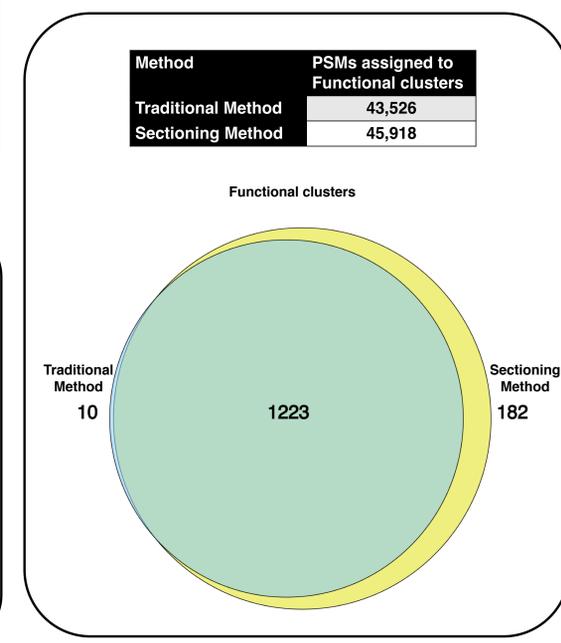


Figure 10: Comparing the number of functional clusters assigned by the PSMs identified by each method using EggNOG in MEGAN6 (Table shows number of PSMs that were assigned to functional cluster)

Figure 11

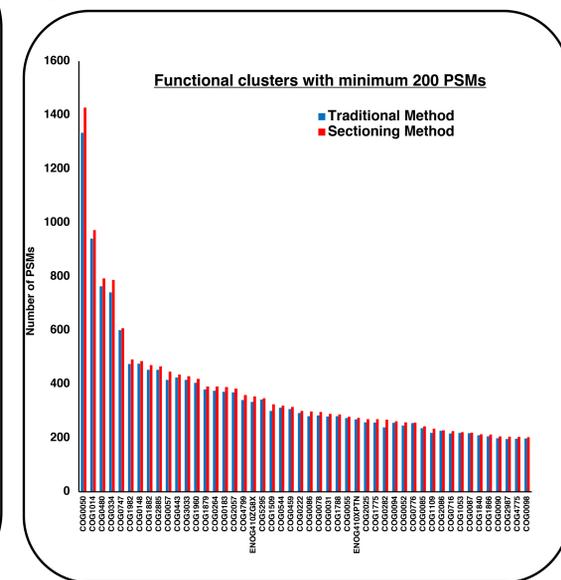


Figure 11: Number of PSMs from each method that were assigned to each functional cluster after using EggNOG in MEGAN6

Results and Discussion

- Multi-stage database searching method has enabled the creation of a reduced database that can be used to match MS/MS data in second step
- We observed 10% increase in the number of PSM identifications in both proteogenomics and metaproteomics database searching
- More variant peptide PSMs were identified to explore in the Proteogenomics dataset
- Comprehensive workflow for proteogenomics (Figure 6) allowed to integrate downstream steps (Figure 5), allowing to check spectral quality and determine the genomic localization of variant peptides.
- Sectioning large database in a metaproteomic study and then using multi-stage two-step identified more number of PSMs
- Most of the genera and functional clusters assigned by PSMs identified by traditional method were also assigned by PSMs identified by the sectioning method.
- All the assigned genera and functional clusters showed improved PSM assignments by using sectioning method

References

- Heydarian et al. (2014) (DOI: 10.4172/jpb.1000302)
- Rudney, J.D. et al. (2015) (DOI: 10.1186/s10168-015-0136-z)
- Jagtap, P. et al. (2013) (DOI: 10.1002/pmic.201200352)
- Woo, S. et al. (2015) (DOI: 10.1021/acs.jproteome.5b00264)
- Sheynkman, G. M. et al. (2014) (DOI: 10.1186/1471-2164-15-703)
- Jagtap, P. et al. (2014) (DOI: 10.1021/pr500812t)

Acknowledgements

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

Traditional Database Searching Method

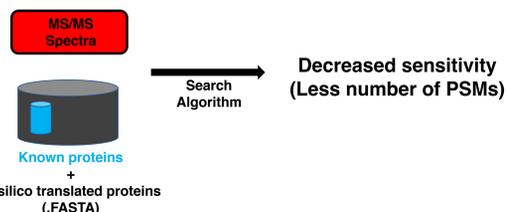


Figure 1: Traditional database searching method where the in-silico translated protein database is appended to the known proteins, increasing the size of the database being searched, thus leading to loss of sensitivity

Increasing Database Size Decreases PSMs

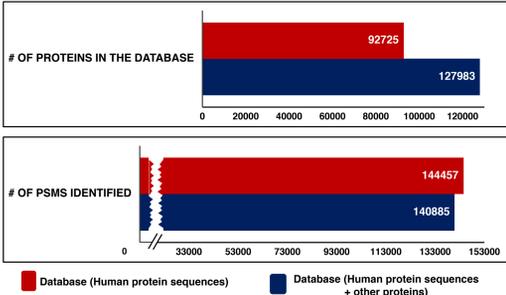


Figure 2: Example showing decrease in number of PSMs with increase in the database size

Method

- In multi-stage method, the MS/MS data is matched to each sequence databases (such as the reference protein sequences, Single Amino acid Variations (SAVs), frameshifts etc.) in a sequential manner.
- MS/MS spectra matched successfully to peptide sequences at each successive stage are removed, and the remaining MS/MS spectra are matched to the next database.
- All the protein sequences, which are identified in the multi-stage method, are used to constitute a reduced database. In the second step, the MS/MS data is matched to this reduced database.
- Following datasets were used to test the methods:
 - Proteogenomics: Samples from mouse early developmental B cells (RNA-Seq and MS/MS) – Heydarian et al. (2014)
 - Metaproteomics: Proteins from dental caries samples grown *in vitro* – Rudney, J.D. et al. (2015)

Figure 3

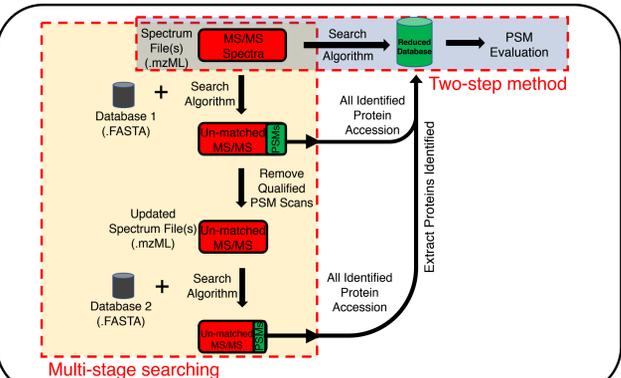


Figure 3: Multi-stage two-step database searching: First, multi-stage searching is performed where the matching of MS/MS data against a database, removing the matched spectra, and matching the un-matched MS/MS spectra against next database. A reduced size database is created using the results from the above multi-stage database searching. The MS/MS spectra are matched to this reduced database in the second step.

Figure 4

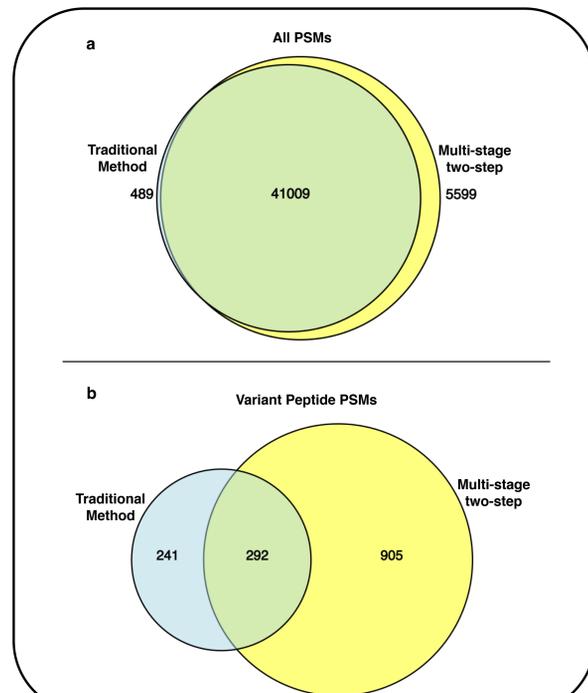


Figure 4: (a) Comparing the number of all qualified PSMs identified by each method. (b) Number of variant peptide PSMs identified by each method and their overlap.

Figure 5

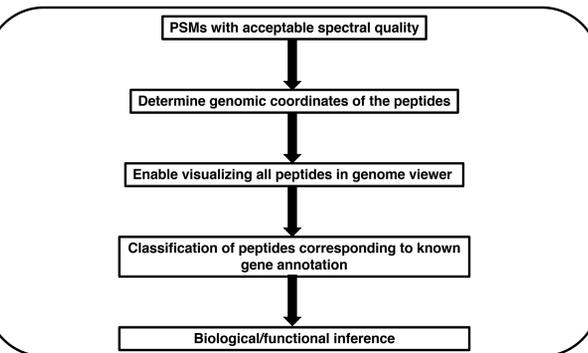


Figure 5: Galaxy-P proteogenomics workflow including downstream analysis steps

Figure 6

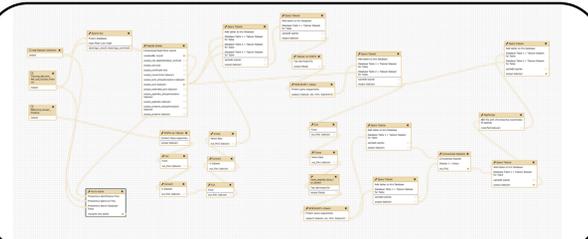


Figure 6: Galaxy-P proteogenomics workflow including downstream analysis steps