De Novo Library Construction for Metaproteomic Analysis of Human Datasets

Rajczewski, A.T.*, Wagner, R.†, Mehta, S.*, Jagtap, P.D.*, Griffin, T.J.*

*Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota †Minnesota Supercomputing Institute, University of Minnesota

Center for Metabolomics and Proteomics
University of Minnesota
Driven to Discover℠

1 Motivation

- **Metaproteomics** uses bottom-up mass spectrometry to gain taxonomic and functional information on microbial communities.
- As with all bottom-up proteomics, metaproteomics requires a FASTA file of anticipated species present
- When interrogating mass spectral data for metaproteomic insights, there are often no metagenomic data available to describe the environment

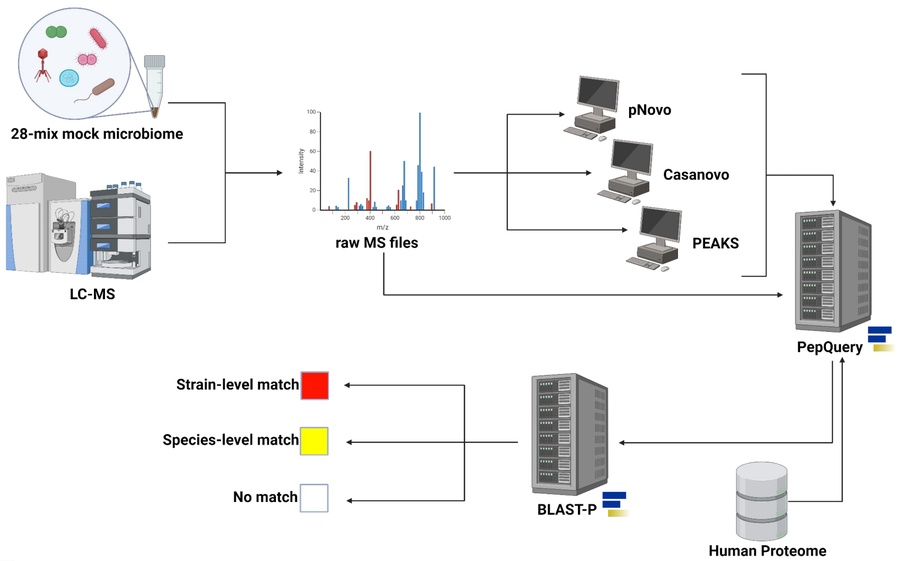
2 Hypothesis

- *De novo* sequencing of mass spectra to yield peptide sequencing has made great strides in recent years
  - Numerous open-source and for-purchase options exist
- ***De novo sequencing, when paired with technologies to remove contaminating sequences and identify contributing species, can be used to build databases for metaproteomic analyses***

3 Samples & Methods

- **Samples**: 1µg replicates of constructed microbial communities
- **LC**: Ultimate 3000 UHPLC in C18 nano mode with a 90-minute gradient
- **MS**: QExactive Orbitrap Quadrupole Hybrid Mass Spectrometer
- **Urinary proteome mass spectrometry data** from Yu *et al*. DOI: 10.7150/thno.16086
- **Casanovo, pNovo, PEAKS:** utilized for *de novo* sequencing
- **PepQuery, BLAST-P,** and **Uniprot** were utilized to refine and construct a database
- **MaxQuant** software suites utilized for quantitative proteomics analysis

4 *De Novo* Sequencing Software Selection

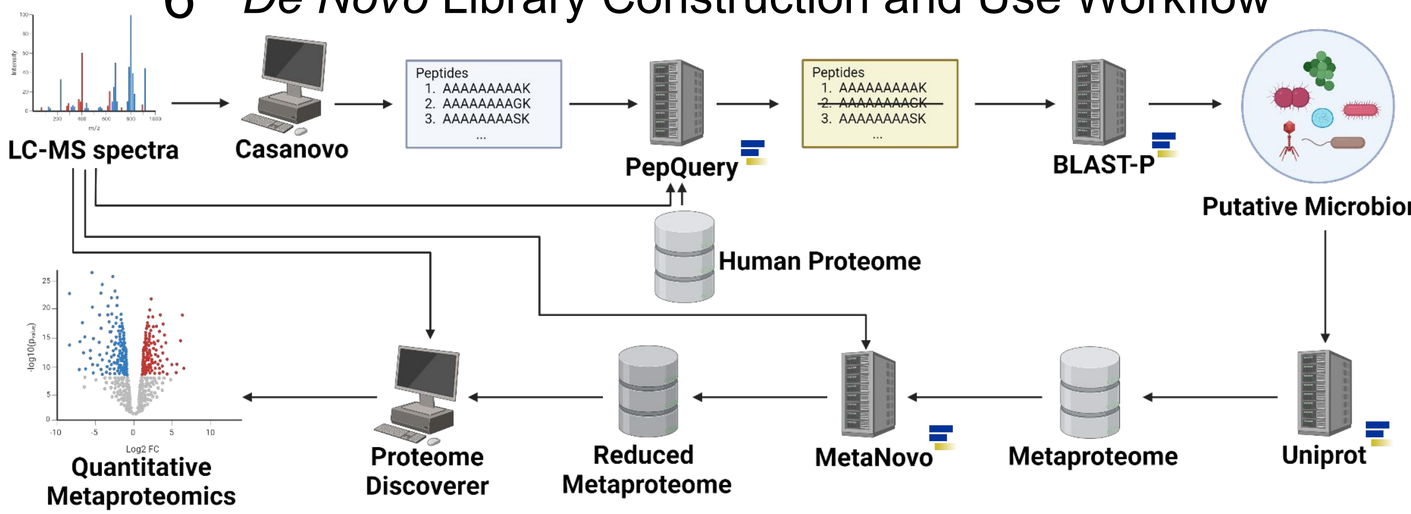


5 Casanovo for *De Novo* Sequencing

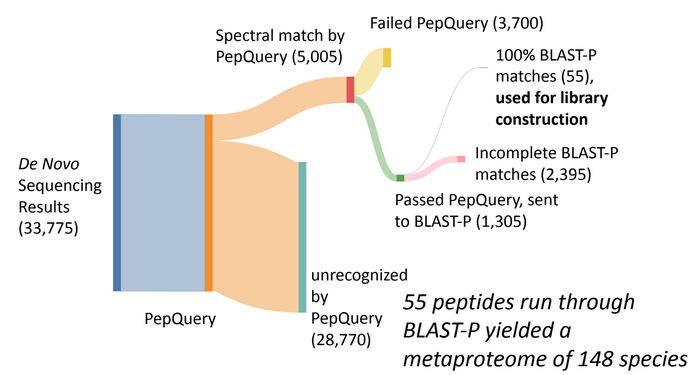
	PEAKS	Casanovo	pNovo	Protein abundance %
Salmonella enterica typhimurium LT2				33.774
Cupriavidus metallireducens CH34				15.519
Stenotrophomonas maltophilia SelTE02				8.021
Pseudomonas fluorescens ATCC 13525, Type strain				6.696
Escherichia coli K12 with Flac+ Plasmid				5.788
Agrobacterium tumefaciens NTL4				5.647
Chlamydomonas reinhardtii				3.996
Rhizobium leguminosarum bv. Viciae				3.172
Pseudomonas denitrificans ATCC 13867				2.871
Staphylococcus aureus subsp. aureus Rosenbach				2.606
Thermus Thermophilus HB27				1.68
Roseobacter sp. AK199				1.596
Chromobacterium violaceum CV026				1.259
Pseudomonas pseudoalcaligenes KF707				1.165
Alteromonas macleodii ATCC 27126				0.954
Desulfovibrio vulgaris Hildenborough				0.946
Paracoccus denitrificans ATCC 17741				0.922
Nitrososphaera viennensis				0.819
Bacillus subtilis 168				0.788
Nitrosomonas ureae Nm10				0.543
Burkholderia xenovorans LB400				0.433
Nitrosospira multiformis ATCC 25196				0.209
Phage M13				0.147
Phage P22 (HT105)				0.106
Phage F0				0.088
Phage ES18 (H1)				0.088
Phage F2				0.084
Nitrosomonas europaeae ATCC 19718				0.082

Strain-level (red) and species-level (yellow) matches of the three de novo sequencing softwares are presented here. Casanovo was selected for use with human datasets due to its high number of matches and potential for integration into Galaxy

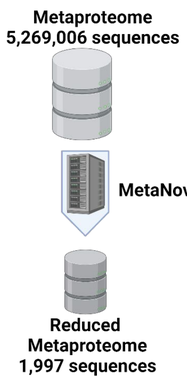
6 *De Novo* Library Construction and Use Workflow

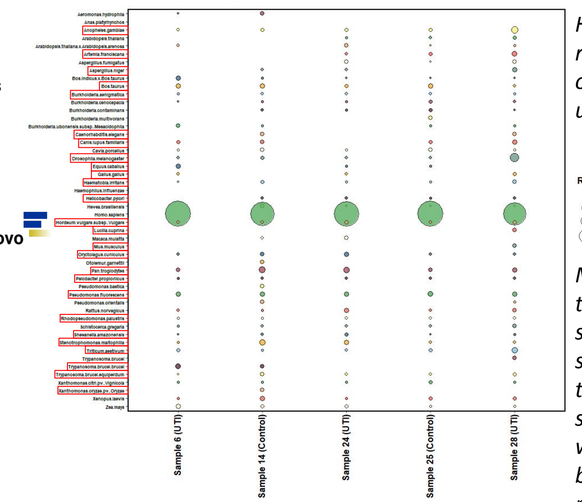


7 *De Novo* Metaproteomic Analysis



55 peptides run through BLAST-P yielded a metaproteome of 148 species





Homo sapiens represent the bulk of the human urinary proteome

MS results suggest the "presence" of several unlikely species, indicating this de novo sequencing workflow may not be suitable for metaproteomics

8 Discussion / Recommendations

- Casanovo was selected for *de novo* library construction
- Of the 33,775 peptides generated by *de novo* sequencing, 55 peptides showed 100% sequence alignment in BLAST-P
- *De novo* sequencing posits unlikely species in the urinary metaproteome
- Other *de novo* sequencing platforms will be tested in future studies
- Future experiments will automate the workflow in the Galaxy suite.

9 Acknowledgements and Funding

This research was supported in part by the National Institutes of Health grant P01 CA138338. Andrew Rajczewski was supported by an NIH biotechnology training grant T32GM008347 from the NIH National Institute of General Medical Sciences. Workflows generated via biorender.com