

EVALUATING CUSTOMIZED DATABASE GENERATION METHODS FOR METAPROTEOMICS ANALYSIS.

Subina Mehta¹, Thomas McGowan¹, James E Johnson¹, Praveen Kumar¹, Ray Sajulga¹, Magnus Arntzen², Francesco Delogu², Marie Crane¹, Peter S.Thuy-Boun³, Dennis W Wolan³, Timothy J Griffin¹, Pratik D Jagtap¹.

1. University of Minnesota, Minneapolis, MN, USA; 2. Norwegian University of Life sciences, Ås, Norway; 3. Scripps Research, La Jolla, CA, USA

INTRODUCTION

- In metaproteomics, the choice of protein sequence search database plays critical role for identification of peptides/proteins from mass spectrometry data .
- Database size and composition presents challenges for optimal identification.
- In this study, we evaluate contemporary database generation software tools used to generate customized, compact and annotated protein sequence search databases.
- In particular, we test different approaches of generating protein sequence FASTA database from metatranscriptomics and metagenomics data.

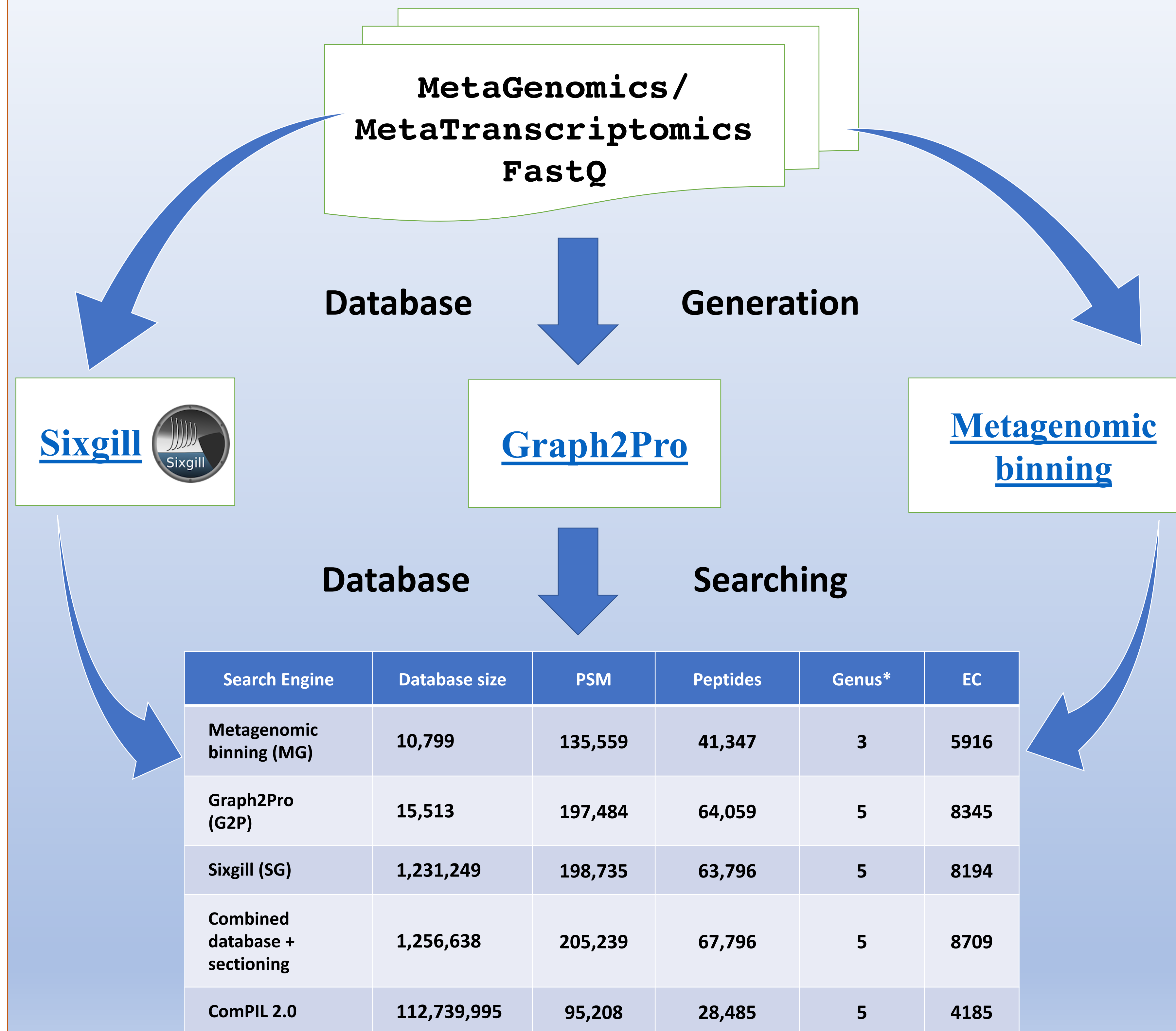
Approaches used:

- Graph2Pro** (metatranscriptomic) - a graph-centric approach that employs the *de Bruijn* graph structure reported by metagenome assembly algorithms to generate a comprehensive database. The assembled graph/putative peptides are compared with the supporting MS/MS spectra to obtain the potential protein sequences.
- Sixgill** (metagenomics) – This method generates protein fragments that are obtained by six frame translation of site-specific metapeptides.
- Metagenomic binning** (metagenomics) – This method involves assembling metagenomic reads in contigs using **metaSPAdes**, the taxonomic binning of contigs was done using **MaxBin2** with a minimum contig length of 5000.
- Sectioning** of the Combined Database (Sixgill + Graph2Pro + Metagenomic Binning)

- Results were compared with **ComPIL 2.0** (for search datasets with unknown composition)
- Taxonomic and Functional annotation was performed to evaluate the outputs.

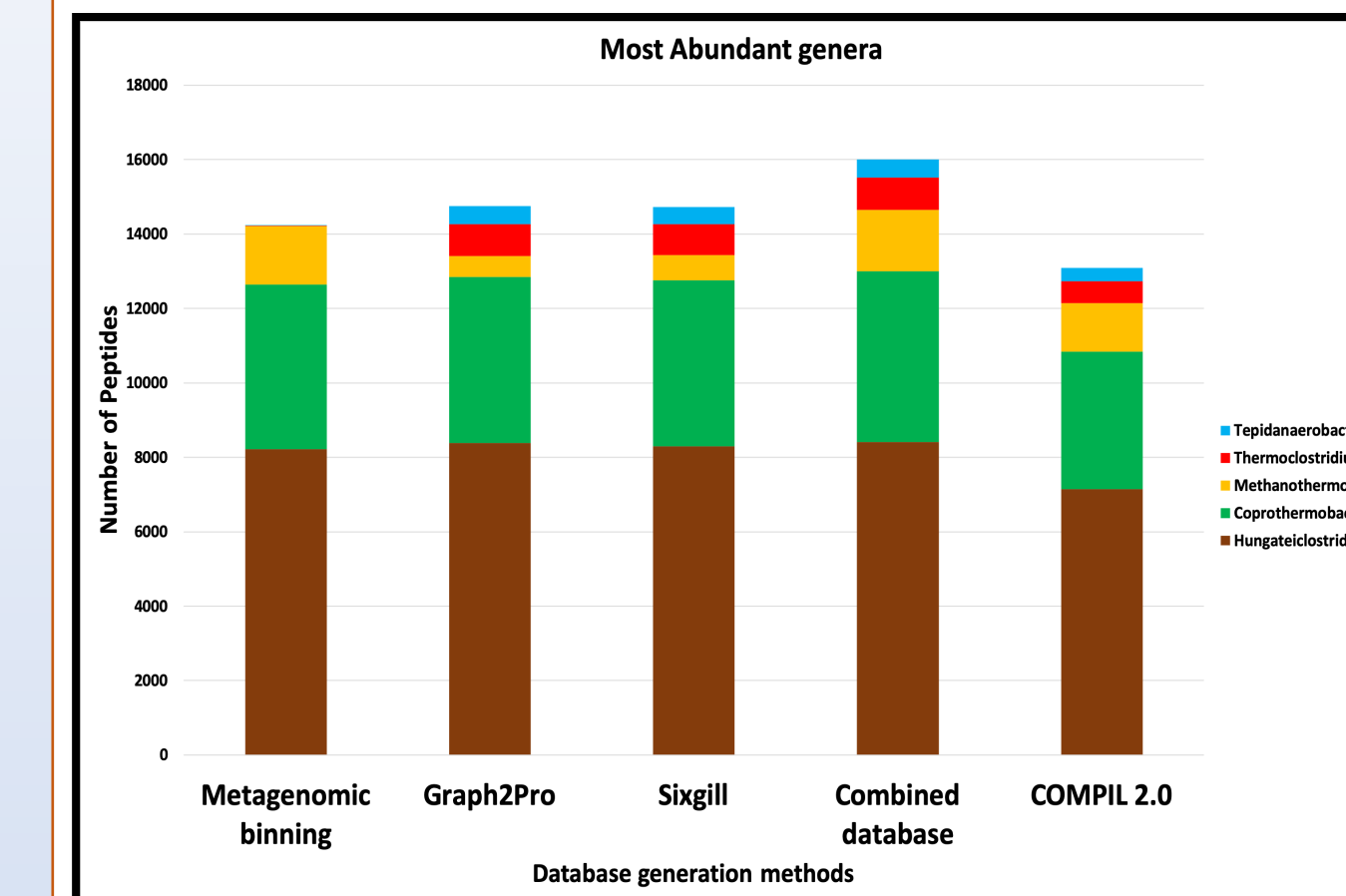
METHOD

- Metatranscriptomics / metagenomics data was obtained from a [time course study](#) of an anaerobic cellulose degrading community.
- Mass Spectrometry raw files were searched against the databases generated using Sixgill, Metagenomic binning, Graph2Pro and the combined database using [SearchGUI/Peptide Shaker](#) and the [sectioning method](#) (for combined database) within the Galaxy platform.
- The PSM (1% FDR), Peptide (1% FDR) outputs generated were then compared with **ComPIL 2.0** (1% FDR).
- For functional and taxonomy annotation of these peptides, we used [Unipept 4.3](#) web interface.

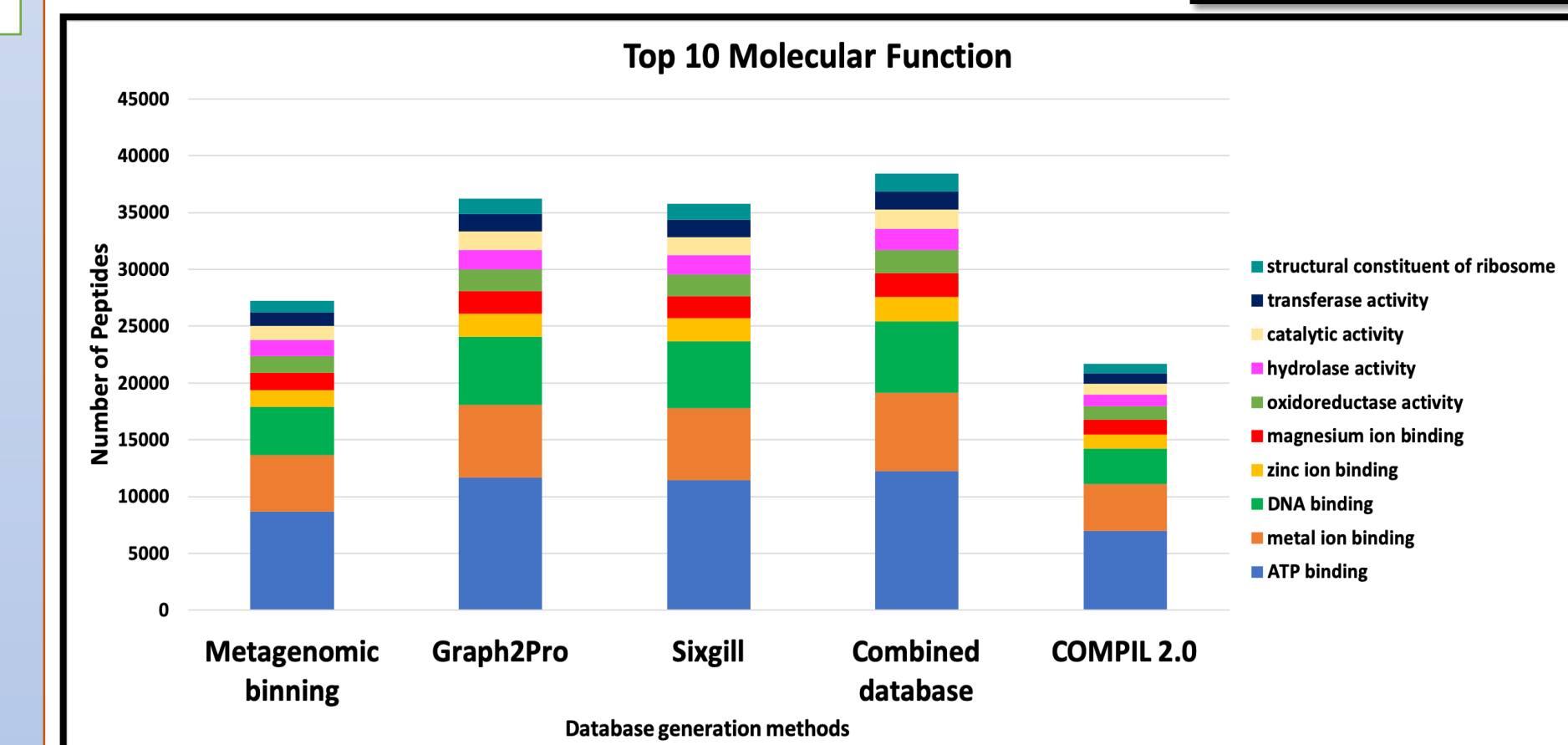


* Genus with 0.5% peptide filter

RESULTS

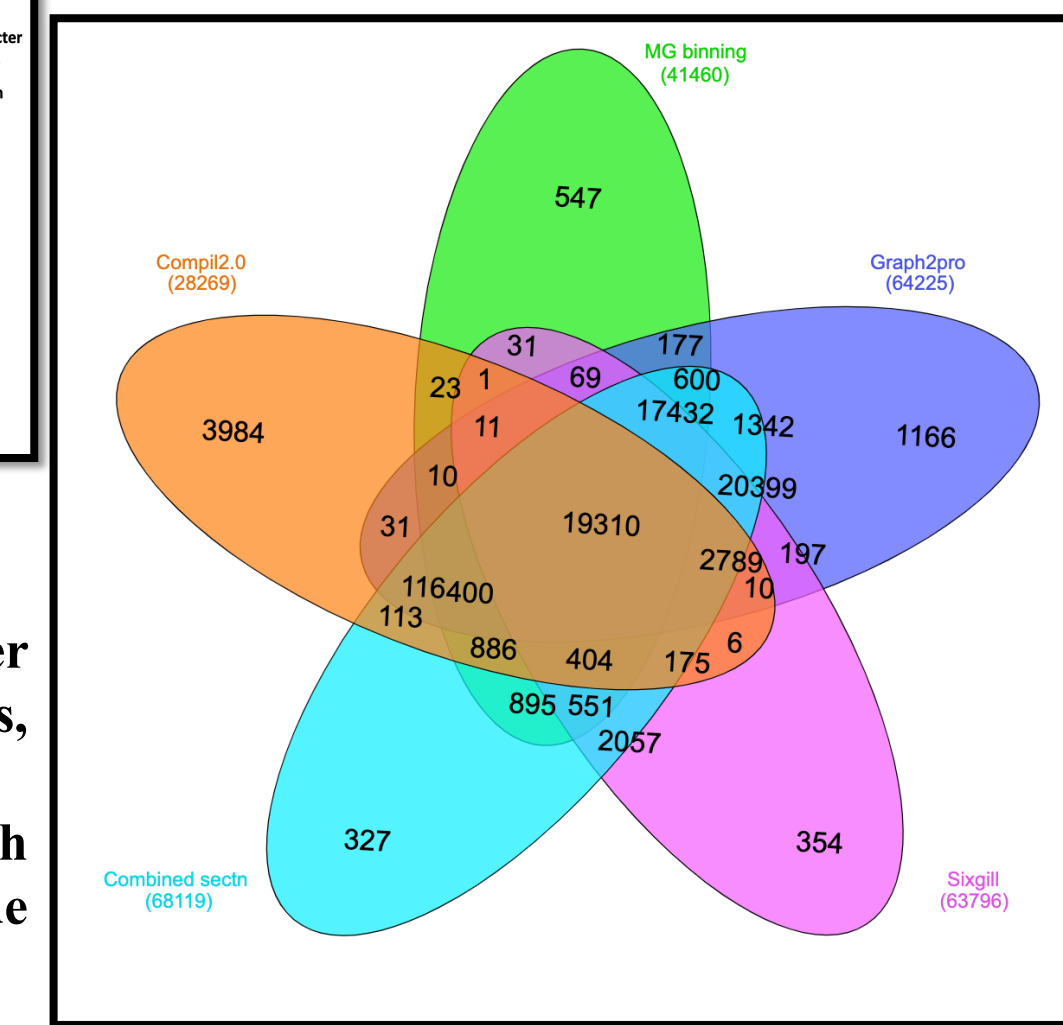


- Peptide overlap between the methods.
- COMPIL 2.0 and Metagenomic binning had lesser peptide identification compared to the others, presumably due to the size of their databases.
- Combined database with sectioning approach provided with maximum number of peptide identifications.



- Top ten molecular function GO terms are plotted here.
- Throughout the methods we see that ATP binding, metal ion binding and DNA binding are the most abundant GO terms.

- This plot represents the most abundant genera obtained from the Unipept after **0.5% filter**.
- All the methods follow a similar trend, however, there are less peptides detected that belong to *Thermoclostridium* and *Tepidanaerobacter* with metagenomic binning.



CONCLUSION

Our primary evaluation shows that the searches with the sectioned combined database provides better identification statistics (such as peptides identified; taxonomy and functional assignments) as compared to other individual approaches.

Funding support from Grant NCI-ITCR grant 1U24CA199347, NSF grant 1458524, 2018 Norwegian Centennial Chair Program Seed Grant. Thanks to the Galaxy Europe for access to software and storage on usegalaxy.eu