Unravelling The Functions of Microbiomes: A Comprehensive Evaluation of Software Tools for Functional Metaproteomics

University of Minnesota
Driven to Discover®

Caleb Easterly[1], Carolin Kolmeder[2], Thilo Muth[3], Bart Mesuere[4], Subina Mehta[1], Praveen Kumar[1], James Johnson[5], Shane Hubler[1], Jaime Huerta-Cepas[6], Bjoern Gruening[7], Michael Riffle[8], Damon May[8], W. Judson Hervey[9], Alessandro Tanca[10], Brook L Nunn[8], Joel Rudney[1], Timothy J. Griffin[1], & Pratik D. Jagtap[1]

[1]Biochem. Mol Biol. Biophysics, University of Minnesota, Minneapolis, MN; [2]University of Helsinki, Helsinki, Finland; [3]Robert Koch Institute, Berlin, Germany; [4]Ghent University, Ghent, Belgium; [5]Minnesota Supercomputing Institute, UMN, Minneapolis, MN; [6]European Molecular Biology Laboratory, Heidelberg, Germany; [7]University of Freiburg, Freiburg, Germany; [8]University of Washington Genome Sciences, Seattle, WA; [9]Naval Research Laboratory, Washington DC; [10]Porto Conte Ricerche Science and Technology Park of Sardinia, Alghero, Italy

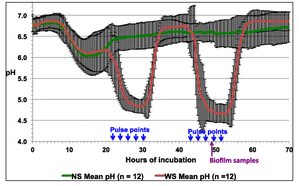
INTRODUCTION

- Metaproteomics research involves large-scale characterization of the entire protein complement of the microbiome. Metaproteomics has the potential to unravel the mechanistic details of microbial interactions with the host/environment by analyzing the microbiome's functional dynamics at the moment of analysis
- Many methods have been developed to determine the functional role of proteins expressed by the microbiome and subsequently shed light on its biological significance. The available software tools differ in emphasis, features, reproducibility, and other characteristics.
- By using a previously published oral microbiome dataset, we explore the following qualitative and quantitative features of several functional analysis software tools (listed below):
  - i) number of functional terms obtained
  - ii) different functional ontologies available
  - iii) ability to leverage quantitative information
  - iv) visualization of datasets
  - v) biological conclusions drawn from data
  - vi) reproducibility, ease of use, and availability.

- **eggNOG mapper**: Huerta-Cepas, et al., 2017
- **MEGAN**: Huson, et al., 2016
- **MetaGOmics**: Riffle, et al., 2017
- **MetaProteomeAnalyzer**: Muth, et al., 2018
- **Unipept**: Mesuere, et al., 2018 (functional analysis version, in beta)

DATA

- Mass spectral data (Rudney, et al., *Microbiome* 2015; PRIDE PXD003151) were acquired from plaque sampled from a patient at high risk for dental caries and grown in biofilm reactor in the presence and absence of sucrose (WS and NS, respectively).
- Previous functional analysis of the data had shown sucrose-induced changes in protein relative abundance patterns for several metabolic pathways
- Mass spectra were searched against the Human Oral Microbiome database (HOMD) to obtain peptide sequences, and spectral counts were calculated for each peptide



Effects of sucrose pulsing on microbiome pH (Rudney, et al., *Microbiome* 2015)

METHODS

We analyzed the data with several functional analysis tools (see tool list in **Introduction**), using standard procedures for each tool and the input files indicated in the **Tool Features** table.
The outputs were compared using several methods:

1) Fold changes (WS over NS) were calculated based on spectral counts and ranked for the output protein, GO term, or orthologous group. The top 5 were compared across the tools to determine the level of consistency.
2) The outputs from each tool were translated to GO terms (see procedure to the right). To determine if any differences were due to more specific GO terms, we mapped the obtained GO terms to the GO generic slim, which contains only 232 high-level terms, compared to the 47,248 GO terms in the full ontology (as of 5/29/2018)

Translation to GO terms

MEGAN
Query eggNOG API with orthologous group IDs

MetaProteomeAnalyzer:
Query UniProt API with protein IDs

Other tools:
GO terms are already provided

QUANTITATIVE RESULTS

Identification Statistics

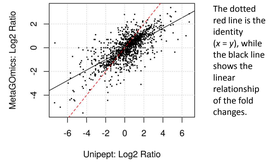
Tool	Native Output			GO term translation			
	Type	# Total	# Significant at FDR < 5%	# Total	# Exclusive to Tool	# Slim Total	# Slim Exclusive
eggNOG mapper	Proteins	18,440	NA	6,155	3070	144	13
MEGAN	eggNOG orthologous groups	1,665	NA	1,450	103	82	0
MetaGOmics	GO terms	3,944	1,958	3,944	645	116	0
MPA	Proteins	23,169	NA	1,102	15	103	0
Unipept	GO terms	2,036	NA	2,036	217	123	0

Comparison of number of identifications obtained by each tool. The "exclusive" columns indicate the number of GO terms that were obtained exclusively by each tool. Note that MetaGOmics is the only tool that performs statistical tests with 2 samples.

GO Term Similarities

Jaccard indices between the sets of GO terms, where 1 indicates identical and 0 disjoint. The Jaccard index is defined as $J(A,B) = |A \cap B|/|A \cup B|$ - that is, the size of the intersection divided by the size of the union. When mapped to the generic GO slim, the sets are more similar. In addition, MEGAN is the least similar to the others.

MetaGOmics and Unipept Fold Change Correlation



The dotted red line is the identity ($x = y$), while the black line shows the linear relationship of the fold changes.

For each GO term identified by both Unipept and MetaGOmics (1625 GO terms, or 80% of Unipept's terms and 41% of MetaGOmics' terms), a fold change (WS over NS) was estimated. The Pearson correlation coefficient was 0.693 ($p < 10^{-16}$). MetaGOmics and Unipept were the only two tools that report spectral counts associated with the same identifiers (GO terms)

Top 5 Upregulated Proteins/GO Terms/Orthologous Groups

Rank	eggNOG mapper (eggNOG proteins)	MEGAN (eggNOG orthologous groups)	MetaGOmics (GO terms)	MetaProteomeAnalyser (UniProtKB proteins)	Unipept (GO terms)
1	Peptidase propeptide and YPED domain	Streptococcal surface antigen repeat	Pyruvate oxidase activity	Ferritin	Peptide deformylase activity
2	Orn lys arg decarboxylase	Acetolactate synthase	Oxidoreductase activity, acting on the aldehyde or oxo group of donors, oxygen as acceptor	Non-heme iron-containing ferritin	PFK-1 activity
3	Glycosyl hydrolase family 70	Catalyzes the condensation of the acetyl group of acetyl-CoA with 3-methyl-2-oxobutanoate to form 3-carboxy-3-hydroxy-4-methylpentanoate	Serine-type endopeptidase inhibitor activity	Clp protease ClpX	D-tagatose 6-phosphate catabolic process
4	DNA protection during starvation protein	Beta-hexosaminidase	Response to wounding	Chaperone protein DnaK	Tagatose-6-phosphate kinase activity
5	Oxidoreductase required for the transfer of electrons from pyruvate to flavodoxin	Cell wall	Poly(ribitol phosphate) teichoic acid metabolic process	ATP-dependent protease ATP-binding subunit	Response to wounding

TOOL FEATURES

	eggNOG Mapper	MEGAN	MetaGOmics	MetaProteomeAnalyzer	Unipept
Inputs	Peptide list	Peptide list, search database, BLAST-P results	Peptides w/spectral counts, search database	Spectrum files (MGF), search database	Peptide list
Outputs	Peptides annotated with protein hits and functional terms	Many options – we used eggNOG orthologous groups with spectral counts	GO terms with fold changes and associated statistical significance	UniProt protein IDs	GO terms with spectral counts
Level of Analysis	Peptide (terms from protein orthologs)	Protein	Peptide	Meta-protein / protein groups	Peptide
Annotation database	eggNOG	NCBI nr	UniProtKB	UniProtKB	UniProtKB
Functional ontologies	COG categories, GO terms, BiGG reactions, KO groups	InterPro2GO, eggNOG orthologous group, SEED and KEGG	GO Terms	EC numbers, protein-level information from UniProt (ontology terms), KEGG, KO groups	GO terms, EC numbers
Comparative Analysis of Multiple Samples	Yes	Yes	Yes (2 samples only)	No	No
Quantitation	Manual (Spectral Counts, MS1 intensities)	Spectral Counts	Spectral Counts	Spectral Counts	Spectral Counts
Functional Visualization	Downstream processing required	Heatmaps, PCA plots, hierarchical cluster analysis, tree diagrams, rarefaction curves	Static GO hierarchy colored by up/downregulation	Interactive bar + pie charts	Interactive treeview of E.C. numbers
Operating System	macOS, Linux	macOS, Linux, Windows	Web	macOS, Linux, Windows	Web
Open Source	Yes	No	Yes	Yes	Yes
Customizability of Analysis	Moderate	High	Low	Low	Low

SUMMARY & CONCLUSIONS

Different tools show very different results with the same data

- The tools returned very different lists of GO terms
  - When mapped to the GO slim, the differences were reduced, suggesting that more specific terms drive the majority of the difference
  - MEGAN was generally the least similar to the other tools
  - eggNOG mapper provided by far the most GO terms
- When the functional objects (protein, GO term, or orthologous groups) were ranked by fold change, the top 5 results provided by each tool show very little overlap
  - The reasons for this are unclear, but may be partially due to the tools' different databases and mapping approaches
- Even for two tools using the same ontology and database (Unipept and MetaGOmics), fairly different results were seen.
  - The fold changes had a modest correlation (0.693), and the GO term lists had a Jaccard index of only 0.37
- In general, it is difficult to compare results due to the use of different ontologies by the different tools

Microbiome functional analysis tools offer a variety of analysis paradigms

- Interactive versus automated
- Different ontologies
- Peptide-centric versus protein-centric
- Quantitative versus qualitative

The tools analyzed here do not support fully quantitative analysis

- Fully quantitative analysis requires the ability to analyze multiple samples and the use of labeled or label-free quantitative values

FUTURE DIRECTIONS

Ultimate goal: Fulfill the functional analysis needs of microbiome/metaproteome researchers

- Explore more fully the reasons for the discrepancies between tools
- Provide a benchmark dataset containing known functions
- Identify ontologies best suited for microbiome studies
- Allow analysis of multiple samples and inter- and intra samples comparison
- Promote and move towards fully quantitative analysis

Reaching this goal requires the collaboration of microbiologists, metaproteomics researchers, and bioinformaticians.

ACKNOWLEDGEMENTS

- Thanks to Minnesota Supercomputing Institute at University of Minnesota for providing essential infrastructure
- This project is supported by National Science Foundation (NSF) grant 1458524 and National Institutes of Health (NIH) grant U24CA199347.
- The opinions and assertions contained herein are those of the authors and are not to be construed as those of the US Navy, the military service at large, or the US Government.