Quantifying Functional Microbiomes: An integrated, quantitative metaproteomics approach reveals connections between taxa, function and protein expression in complex microbiomes

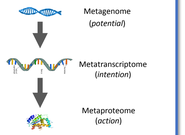
Pratik Jagtap[1], Caleb Easterly[1], Nadia Szeinbaum[2], Lee S. Parsons[1], Andrea Argentini[3,4], Thomas McGowan[1], Bjoern Gruening[5], Alessandro Tanca[6], Shane Hubler[1], Carolin Kolmeder[7], Subina Mehta[1], Bart Mesuere[3,4], James E. Johnson[1], Praveen Kumar[1], Lennart Martens[3,4], Jennifer Glass[2], Joel Rudney[1], Brook L. Nunn[8], Timothy J. Griffin[1]

[1]University of Minnesota, Minneapolis, MN; [2] Georgia Institute of Technology, Atlanta, GA;[3] Ghent University, Ghent, Belgium; [4]VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium; [5]University of Freiburg, Freiburg, Germany; [6]Porto Conte Ricerche, Science and Technology Park of Sardinia, Alghero, Italy; [7]University of Helsinki, Helsinki, Finland; [8]University of Washington, Seattle, WA

INTRODUCTION

- Microbiome research offers promising insights into microbial contributions to human and environmental systems.
- Metaproteomics enables functional analysis of the microbiota by analyzing its proteins
- Many metaproteomics studies are qualitative (detection) or semi-quantitative (using spectral counts). However, precursor (MS1) intensity has been shown to offer a wider dynamic range than spectral counts
- In addition, the interplay between protein function and organismal source has often been overlooked
- We attempt to answer 4 questions for a given microbiome using quantitative information:
  - What organisms are present, and how abundant are they?
  - What processes are the organisms carrying out?
  - Which organisms contribute to each function?
  - Can we characterize proteins of unknown function?
- Our methods can be used for both exploratory and hypothesis-driven data analysis.

Action: Why Metaproteomics is Important



Proteins are dynamic and rapidly turned over. Proteomics reveals the molecular machinery in action at the time of harvest, when experimental conditions are known.

METHODS

We propose methods to analyze metaproteomics data on the level of function, taxonomy, and the function-taxonomy interaction.

1) Analysis of function
- a) Determine total MS1 intensity associated with an assignment of interest (ex: GO term, COG category, taxon) - see **Figures 1** and **2**, below
- b) Normalize intensity to root node (ex: *biological process*)
- c) Calculate intensity ratio between experimental conditions, and perform t-tests to determine statistical significance

2) Analysis of taxonomic distribution
- a) Same as function

3) Analysis of the function-taxonomy interaction
- a) Group intensities by COG category and lowest common ancestor
- b) Use PECA (Suomi, et al., *J Proteome Res* 2015) to calculate fold changes and determine statistical significance

Function

- Microbiome processes can be understood through function classification of proteins
  - Protein function is described using the language of ontologies
    - Examples:
      - Gene Ontology (GO)
      - Enzyme Classification (EC) numbers
      - Clusters of Orthologous Groups (COG) categories
  - Several ontologies are hierarchical
  - For GO terms, we use the approach from metaGOmics (Riffle, et al., 2017) – see **Figure 1**
  - In addition, the option is given to use the full GO or the generic slim GO, which is a higher-level view of function

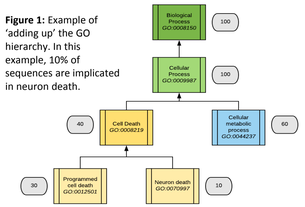


Figure 1: Example of 'adding up' the GO hierarchy. In this example, 10% of sequences are implicated in neuron death.

Taxonomy

- Taxonomy is also a hierarchy (refer to Figure 2, below)
  - *Clostridiales* (order) implies *Clostridia* (class), *Firmicutes* (phylum), and *Bacteria* (superkingdom)
  - Example:
    - a peptide sequence is mapped to a protein produced by all members of the *Clostridiales* order
    - all organisms in *Clostridiales* also belong to the *Clostridia* class and so on
    - therefore, the peptide 'belongs' to (or comes from) the *Clostridia* class, etc.
- Thus, to get the relative abundance of a given taxon, we sum the intensities from all peptides that are mapped to that taxon (and all children of that taxon)

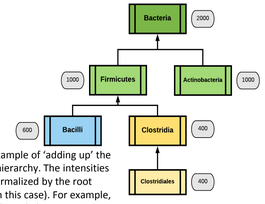


Figure 2: Example of 'adding up' the taxonomy hierarchy. The intensities are then normalized by the root (*Bacteria*, in this case). For example, the order *Clostridiales* contains 20% of all sequences.

Function/Taxonomy Interaction

- For analysing function and taxonomy together, we use COG categories, a high-level, non-hierarchical view of function.
- In addition, we use the lowest common ancestor of the taxonomic assignment, and do not propagate up the taxonomic hierarchy.
- This allows for simple grouping of intensities by function and taxonomy together, and the use of PECA (Suomi, et al., *J Proteome Res* 2015)

Unknowns

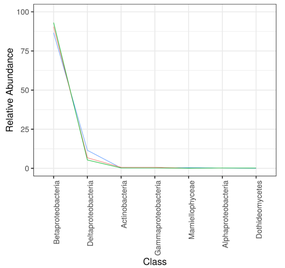
- In most metaproteomics experiments, a significant number of peptides remain unassigned to any taxon or function
- Methods of working with unknowns include:
  - Identifying peptides that correspond to proteins of unknown function that are differentially expressed between experimental conditions
  - Predict protein function based on co-expression with proteins of known function (methods under development)

EXPLORATORY ANALYSIS: IRON-RICH ANOXIC LAKE

Dataset

- Sediment samples from Lake Matano, Indonesia, were incubated in three bioreactors (Sz9, Sz10, and Sz11). Three samples (technical replicates) were taken from each bioreactor.
- The search database was generated using metagenomic data from the bioreactors.
- Sz9 and Sz10 were amended with methane to determine whether anaerobic methane would be enriched for and result in differences in metaproteomic expression
- Of the 18,691 quantified peptides, 42.3% were mapped to a taxon by Unipept and 50.2% were annotated with GO terms by eggNOG mapper
  - The few peptides mapped to UniProt by Unipept and to GO terms by eggNOG mapper indicates that many of the peptides were not present in the UniProt/eggNOG database, which suggests that the taxa and functions in the sample are relatively novel

Results: Taxonomy

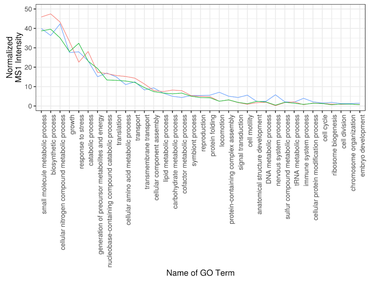


Summed intensities of the peptide sequences specific to a taxonomic class or lower attributed to betaproteobacteria comprised more than 90% of the total intensity of all class-specific sequences.

Preliminary data (not shown) indicate that the *Betaproteobacteria* peptides are produced by a novel species.

Taxonomic distribution at the class level was very similar across bioreactors, consistent with absence of oxidation products of methane (not shown).

Results: Function



Total peptide intensity assigned to terms in the biological process (BP) ontology, divided by the total BP intensity for that sample.

As shown, the BP ontology is dominated by growth and metabolic processes, including nitrogen compound metabolism and biosynthetic processes.

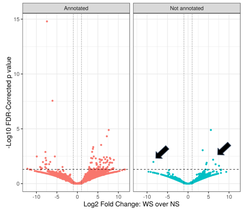
We did not find any noticeable differences between bioreactors, which is consistent with the absence of methane oxidation in the samples.

STATISTICAL ANALYSIS: ORAL DYSBIOSIS

Dataset

- Mass spectral data were acquired from plaque samples from three subjects at high risk for dental caries grown in biofilm reactor in the presence (With Sucrose, or WS) and absence of sucrose (No Sucrose, or NS).
- Mass spectra were searched against the Human Oral Microbiome database (HOMD) to identify microbial peptides.
- Quantitation, functional annotation, and taxonomic assignment were performed as in the workflow below (quantitation and database search performed by MaxQuant)

Results: Proteins of Unknown Function

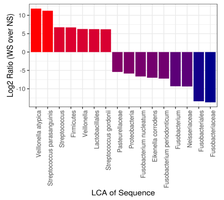


Volcano plot of the protein intensity fold change (WS over NS)
- Most of the differentially expressed proteins are annotated by eggNOG with either KEGG orthologous groups, GO terms, or COG categories (excluding "R" and "S", general functional categorization or unknown)
- Eight differentially expressed proteins have no known function (no KEGG, GO, or COG annotation; indicated with arrows in above plot)

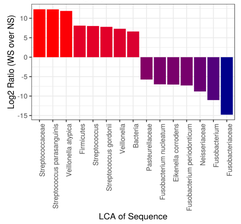
Results: Function-Taxonomy Interaction

The up- or down-regulation of a given function-taxon combination is shown for two COG categories.

Carbohydrate metabolism (COG category G)

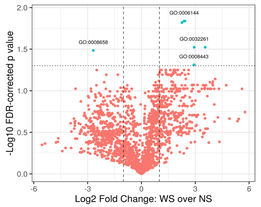


Post-translational modifications, protein turnover, and chaperones (COG category O)



When considering carbohydrate metabolism, the *Firmicutes* phylum and its members tend to participate more in the WS condition, while the *Fusobacteria* phylum tends to participate less in the WS condition. Similar effects are seen with COG category O.

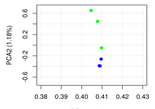
Results: Function



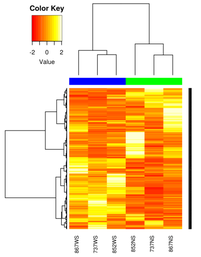
Volcano plot of regulation of GO terms. The dotted lines indicate a fold change cutoff of 2 and a *p*-value cutoff of 0.05

Term	Ontology	Log2 FC
Anti-sigma factor antagonist activity	MF	3.6
transcription factor activity	MF	3.6
Purine nucleotide salvage	BP	2.9
6-phosphofructokinase activity	MF	2.9
nucleotide salvage	BP	2.9
phosphofructokinase activity	MF	2.9
Purine-containing compoud salvage	BP	2.4
GDP binding	MF	2.3
penicillin binding	MF	-2.7

Differentially expressed GO terms. BP = biological process; MF = molecular function.

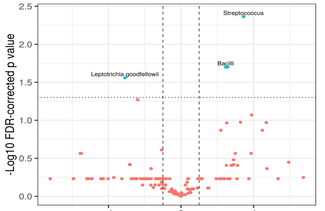


PCA plot of functional distribution. Green indicates NS and blue indicates WS. The sample groups are separated in principal coordinate space to an extent comparable with the taxonomic analysis. More variation is seen in NS than in WS.

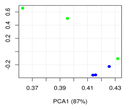


Heatmap of GO term functional distribution. The samples clearly separate into no sucrose (NS, green) and with sucrose (WS, blue) conditions.

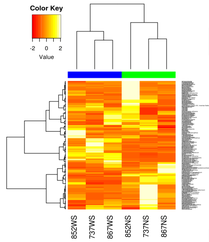
Results: Taxonomy



Volcano plot of differentially expressed taxa. The *Streptococcus* genus, the *Bacilli* class, and the *Lactobacillales* order are higher abundance in WS, while *Leptotricia goodfellowii* is downregulated in WS.

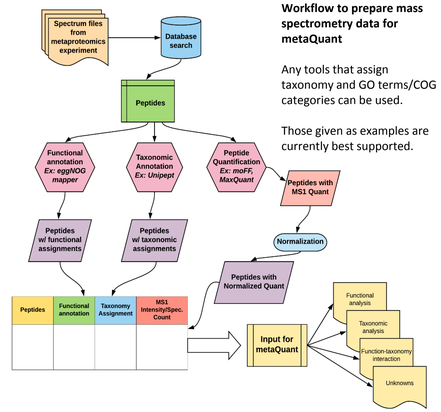


PCA plot of taxonomic distribution. Green indicates NS and blue indicates WS. More variation is seen in the NS group.



Heatmap of taxonomic distribution. The samples (columns) are clustered into with sucrose (WS, blue), and no sucrose (NS, green) conditions. Patterns are less consistent than in the function heatmap.

UPSTREAM ANALYSIS: FUNCTIONAL AND TAXONOMIC ASSIGNMENT



Workflow to prepare mass spectrometry data for metaQuant

Any tools that assign taxonomy and GO terms/COG categories can be used.

Those given as examples are currently best supported.

CONCLUSIONS

- Current microbiome functional and taxonomic analysis methods do not leverage full quantitative information
- The methods of metaQuant allow for understanding the interplay between the taxonomic distribution of microbiota and the functional roles played by each member of the community, critical to fully understand microbiome processes
- More work is necessary to determine boundaries between exploratory and experimental metaproteomics, and appropriate levels of granularity for each
  - more specific terms (e.g., full GO hierarchy) versus more general terms (e.g., COG terms)
- Significant numbers of peptides are not present in reference databases, such as eggNOG (for eggNOG mapper) and UniProtKB (for Unipept). Methods to analyze un-mapped peptides could improve analysis considerably

FUTURE DIRECTIONS

- Add support for E.C. numbers, which are in a hierarchical structure similar to, though simpler than, the GO terms
- Distribute metaQuant via Bioconda and implement metaQuant in Galaxy
- Expand and verify statistical methods
- Explore methods for inferring function of proteins of unknown function
- Implement methods for understanding whether taxonomy or function differs more across experimental conditions
- Develop interactive viewer application

ACKNOWLEDGEMENTS

This project is supported by National Science Foundation (NSF) grant 1458524 and National Institutes of Health (NIH) grant U24CA199347.
We would like to acknowledge the Minnesota Supercomputing Institute (MSI) for maintaining the Galaxy-P server.