Bioinformatics, 36(14), 2020, 4171–4179
doi: 10.1093/bioinformatics/btaa289
Advance Access Publication Date: 5 May 2020
Original Paper

Systems biology

pepFunk: a tool for peptide-centric functional analysis of metaproteomic human gut microbiome studies

Caitlin M. A. Simopoulos [1,2], Zhibin Ning[1,2], Xu Zhang[1,2], Leyuan Li[1,2], Krystal Walker[1,2], Mathieu Lavallée-Adam[1] and Daniel Figeys[1,2,3,*]

[1]Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, ON K1H 8M5, Canada, [2]Faculty of Medicine, SIMM-University of Ottawa Joint Research Center in Systems and Personalized Pharmacology, University of Ottawa, Ottawa, ON K1H 8M5, Canada and [3]Canadian Institute for Advanced Research, Toronto, ON M5G 1M1, Canada

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on November 29, 2019; revised on March 20, 2020; editorial decision on April 24, 2020; accepted on April 27, 2020

Abstract

Motivation: Enzymatic digestion of proteins before mass spectrometry analysis is a key process in metaproteomic workflows. Canonical metaproteomic data processing pipelines typically involve matching spectra produced by the mass spectrometer to a theoretical spectra database, followed by matching the identified peptides back to parent-proteins. However, the nature of enzymatic digestion produces peptides that can be found in multiple proteins due to conservation or chance, presenting difficulties with protein and functional assignment.

Results: To combat this challenge, we developed pepFunk, a peptide-centric metaproteomic workflow focused on the analysis of human gut microbiome samples. Our workflow includes a curated peptide database annotated with Kyoto Encyclopedia of Genes and Genomes (KEGG) terms and a gene set variation analysis-inspired pathway enrichment adapted for peptide-level data. Analysis using our peptide-centric workflow is fast and highly correlated to a protein-centric analysis, and can identify more enriched KEGG pathways than analysis using protein-level data. Our workflow is open source and available as a web application or source code to be run locally.

Availability and implementation: pepFunk is available online as a web application at https://shiny.imetalab.ca/pepFunk/ with open-source code available from https://github.com/northomics/pepFunk.

Contact: dfigeys@uottawa.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Metaproteomics, the study of proteins from an environmental sample, is used to examine the dynamics and composition of microbial communities in complex environments including human and animal microbiomes (Cheng *et al.*, 2018; Moon *et al.*, 2018), soil (Starke *et al.*, 2019) and water samples (Mikan *et al.*, 2020). Understanding the microbial dynamics and functionality of the human gut microbiome is particularly of interest due to its association with human disease as observed in immune-system-associated diseases, such as inflammatory bowel disease (IBD) (Morgan *et al.*, 2012; Zhang *et al.*, 2018b), asthma (Arrieta *et al.*, 2015) and multiple sclerosis (Jangi *et al.*, 2016), metabolic disorders, such as obesity and type-II diabetes (Sonnenburg and Bäckhed, 2016) and cardiovascular disease (Tang *et al.*, 2017). Studies have also demonstrated that the presence of the 'gut-brain' axis can mean that gut microbes are capable of influencing, or are at least linked to, one's mental health, with evidence even suggesting that modulation of microbiota can have therapeutic effects in anxiety and depression (Dash *et al.*, 2015).

Although the term 'proteomics' implies that proteomic data inherently consist of protein-level information, an important step in most proteomic workflows is to enzymatically digest extracted proteins into smaller peptide fragments before mass spectrometry (MS) sequencing (Hettich *et al.*, 2013). To facilitate the analysis, peptides are then separated and analyzed, often by liquid chromatography coupled with tandem MS. Raw spectra produced by MS/MS are computationally matched with predicted spectra of peptide sequences by database search. These matched peptides are then assigned to proteins. However, due to the nature of enzymatic digestion, the same peptide sequence can belong to multiple proteins, and it is difficult to determine the correct parent-protein of these redundant peptides (Nesvizhskii and Aebersold, 2005). Nesvizhskii and Aebersold (2005) deemed this challenge the Protein Inference Problem, which is further exacerbated in metaproteomics experiments due to the presence of multiple microbial strains and species that can include additional redundant peptides due to protein sequence conservation. Nonetheless, computational workflows for

proteomic research typically use proteins identified from peptide sequences for quantitative and functional enrichment studies although redundant peptides can impede accurate and confident identification of proteins from MS/MS data.

Ning *et al.* (2016) describe the uncertainty of peptide-to-protein assignment as 'information degeneration'. This information loss stems from the methods that researchers have previously used to mitigate the ambiguity of peptide-to-protein assignment. For example, the Occam's razor principle relies on discarding proteins without unique peptides, and often can only identify protein groups (Serang and Noble, 2012). Alternatively, Muth *et al.* (2015) have introduced the concept of a 'meta-protein', where proteins are grouped by amino acid sequence or shared peptides. Although methods have been introduced to combat the Protein Inference Problem, methods for a peptide-centric metaproteomic workflow have also been implemented to circumvent information loss. Notably, UniPept is a gene ontology (GO) term-focused functional analysis tool that was released as both a web application and local software tool (Gurdeep Singh *et al.*, 2019). UniPept uses a large GO term functional database consisting of tryptic peptides of proteins found in the UniProt Knowledgebase. However, GO term annotations are organized in a directed acyclic graph with semantic relationships between terms, causing challenges for functional enrichment analyses, such as unclear hierarchies and dependencies (Gaudet and Dessimoz, 2016). To manage this challenge, Riffle *et al.* (2018) created MetaGOmics, a peptide-centric GO term based enrichment tool that creates directly acyclic graphs for GO terms associated to identified peptides. Despite the computational progress, these previously mentioned tools have made for peptide-centric metaproteomic workflows, the tools all use complex GO term annotation and are not specifically created for gut microbiome studies.

In this work, we introduce a novel peptide-centric workflow for metaproteomics data for gut microbiome experiments. To facilitate this work, we created a Kyoto Encyclopedia of Genes and Genomes (KEGG)-to-peptide functional database, a functional enrichment workflow and an interactive web application companion tool for gut microbiome metaproteomic studies. We created a peptide-function database consisting of *in silico* digested peptides from the Integrated human gut microbial Gene Catalog (IGC) database. We reduced the size of our peptide database by focusing on the most empirically identified peptides in raw MS/MS data for improved computational speed. We used annotation from UniProt Reference Clusters 90 (UniRef90) sequence clusters to functionally annotate the gut microbiome peptide database with KEGG terms. We created a peptide-centric functional enrichment workflow by adapting gene set variation analysis (GSVA) for peptide-level data (Hänzelmann *et al.*, 2013). We found that the results from our peptide-centric workflow correlated with results from a protein-centric workflow suggesting that the peptide workflow is suitable and comparable to a more canonical approach to metaproteomics data analysis. Additionally, our peptide-centric workflow was able to identify more enriched KEGG pathways than when using protein-level data. Finally, we packaged our peptide-centric data analysis pipeline into a user-friendly web application intended to be used as a companion tool to MetaLab and iMetaLab (Cheng *et al.*, 2017; Liao *et al.*, 2018), and released the source code to allow local data analysis for experienced computational users.

2 Materials and methods

2.1 KEGG core peptide database construction

We used the IGC protein dataset (https://db.cngb.org/microbiome/genecatalog/genecatalog_human/; Li *et al.*, 2014) to create our KEGG-peptide functional database. We first annotated the IGC protein dataset by searching for sequence identity in UniRef90 sequence clusters (Suzek *et al.*, 2015). Sequence alignment was computed using Diamond blastp (Buchfink *et al.*, 2015) and command line options –sensitive -e 0.1–top 5 -f 6 qseqid qlen sseqid slen evalye length mident, where –sensitive gave us a search with a higher sensitivity, -e 0.1 allowed for a maximum *E*-value of 0.1, –top 5 produced a list of the top five hits and -f let us customize the output file. We considered a single protein match to a cluster sequence in UniRef90 as the smallest *E*-value representing the best match for functional identification. Notably, 99.5% of protein matches have *E*-values <0.0001. Each protein in the IGC dataset was then annotated with KEGG terms using the annotation associated to UniRef90 protein matches. We completed an *in silico* trypsin digestion of the IGC protein dataset using a Python script (https://github.com/northomics/bin/blob/master/trypsin.py) that considered digestion at lysine and arginine except if followed by a proline, and up to two missed enzymatic cleavages. Computed peptides inherited the KEGG functional annotation of parent-proteins. In the case of redundant peptides, or peptide sequences that are found in multiple proteins, the union of all identified KEGG annotations for all parent-proteins were considered (Fig. 1A). In other words, we did not discard any putative functional annotation in redundant peptides and instead multiple KEGG annotations were accounted for in functional enrichment analysis by intensity weighting (Fig. 1B). After *in silico* trypsin digestion, our IGC database of 9 878 647 proteins consisted of 603 457 781 total peptides and 414 419 478 unique peptide sequences. For computational speed, we reduced this database size to peptides frequently matched in human gut microbiome studies to 469 393 unique peptides that were identified from 500 in house, raw MS/MS files. Of these unique peptides, 224 836 (47.9%) have KEGG annotation. Conversely, the IGC protein database has 2 109 127 (21.4%) proteins with KEGG annotation. The reduced database is made available as Supplementary File S1.

2.2 Peptide set and protein set variation analysis

We adapted the GSVA method (Hänzelmann *et al.*, 2013) for use with peptide intensity levels in metaproteomic experiments rather than gene expression estimation from RNA sequencing or microarray experiments. Peptide intensities were corrected by sample-specific size factors to normalize inter-sample variability. We calculated size factors using the DESeq2 R package (Love *et al.*, 2014), which consisted of median ratios of peptide intensities to the geometric mean of each peptide in the entire experiment. Peptide intensities were then divided by their corresponding sample-specific size factor. We removed peptides with intensities missing in over half the samples in each tested condition as a preprocessing filtering step for missing data. We then created 'peptide sets' for our adapted GSVA analysis consisting of peptides annotated in each KEGG pathway for a total of 229 peptide sets. For our protein-level analysis, we also created protein sets from protein groups annotated in each KEGG pathway. Peptides with multiple KEGG annotation terms were included in all appropriate peptide sets by intensity weighting while considering the frequency of KEGG term annotation as explained in Equation (1) and Figure 1B:

$$i_{pk} = x_p(w_{pk}) \tag{1}$$

$$w_{pk} = \frac{n_{pk}}{n_p},$$

where i = adjusted intensity, x = measured peptide intensity, w = weight adjustment, p = peptide, k = KEGG term and n = number of KEGG terms.

We completed a peptide set variation analysis using the GSVA R package (Hänzelmann *et al.*, 2013). We used a Gaussian cumulative distribution function for kernel estimation of each peptide's intensity. To reduce noise, we only considered peptide sets with a minimum size of 10 peptides. The weighted peptide intensities (w_{pk}) were further transformed by $\log_2$. Briefly, GSVA scores, a type of enrichment score, were calculated from ranking peptides in each set and calculating a Kolmogorov–Smirnov-like random walk statistic from these ranked peptide sets. Significant differences in GSVA scores of peptide sets between tested conditions were identified using lmFit(), a least squares linear model, and eBayes(), empirical Bayes moderation, from the limma R package (Phipson *et al.*, 2016) while considering multiple

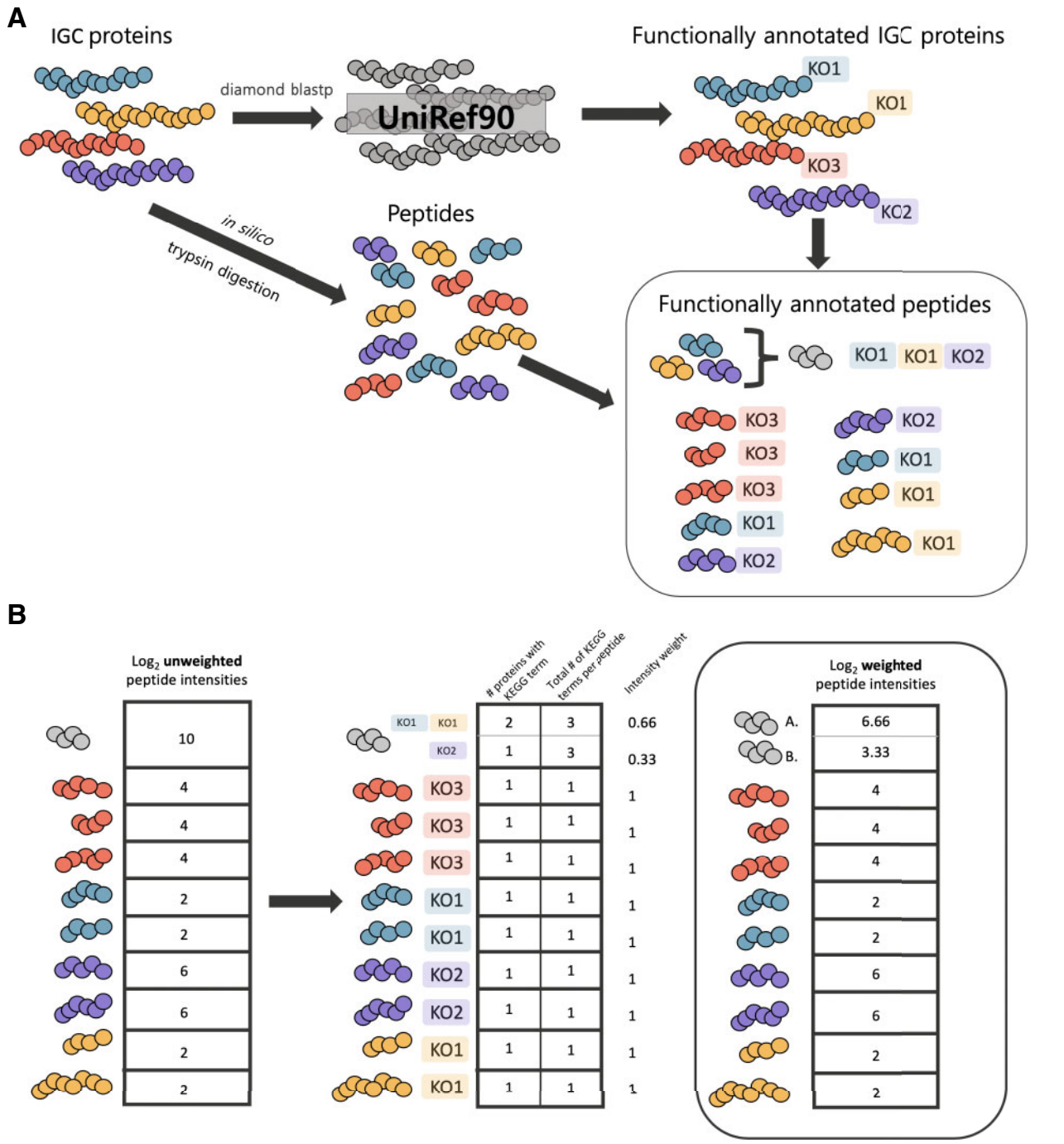

Fig. 1. Illustrated schematic of the peptide database creation and intensity weighting methodology. (**A**) Peptide database protocol. We first used diamond blastp to align IGC proteins to the UniRef90 gene cluster database for annotation of IGC proteins with KEGG terms. In parallel, we completed an *in silico* trypsin digestion of IGC proteins into tryptic peptides that then inherited KEGG annotations from their parent-proteins. Notably, redundant peptides inherited all possible KEGG annotations. In this example, the redundant peptide (gray) can be found in three proteins. (**B**) Intensity weighting of redundant peptides. The intensity of the gray redundant peptide is weighted by our confidence in the peptide's KEGG annotation. The redundant peptide is part of two proteins annotated with KO1 and one protein annotated with KO2, therefore, we weight the peptide's intensity for KO1 and KO2 by 0.66 (2/3) and 0.33 (1/3), respectively. The weighted intensities are then be used in our modified GSVA pipeline where the weighted intensities are associated to the appropriate peptide set. (Color version of this figure is available at *Bioinformatics* online.)

hypothesis testing by adjusting *P*-values using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995).

We compared the results of the peptide set variation analysis workflow to that of a similar workflow centered on protein-level intensity data. The methods remained the same, except we used protein group label-free quantitation (LFQ) intensity values provided by the MetaLab workflow. The protein LFQ values were also normalized by $\log_2$ transformations and then subjected to the same workflow as the peptide intensities including weighting of all protein group intensities with multiple KEGG terms.

2.3 R shiny app construction

We created pepFunk, an R shiny application for our peptide-centric functional enrichment workflow. pepFunk was written in the R programming language v3.4.4 (R Core Team, 2019) and is dependent on the R packages: shiny (Chang *et al.*, 2019), shinydashboard (Chang and Borges Ribeiro, 2018), shinyWidgets (Perrier *et al.*, 2019), DT (Xie *et al.*, 2019) for application building, rhandsontable (Owen, 2018), reshape2 (Wickham, 2007), tidyverse, plyr (Wickham, 2011) for data manipulation, colourpicker (Attali, 2017), tidyverse (Wickham, 2017) and plotly (Sievert, 2018) for custom plotting, DESeq2 (Love *et al.*, 2014), GSVA (Hänzelmann *et al.*, 2013) and limma (Ritchie *et al.*, 2015) for data analysis, ggdendro (de Vries and Ripley, 2016) and dendextend (Galili, 2015) for dendrogram plotting and LaCroixColoR (Bjork, 2019) for color palettes. Dataset 1 is provided as sample data within the app, and was also deposited to the ProteomeXchange Consortium (Deutsch *et al.*, 2017) via the PRIDE (Perez-Riverol *et al.*, 2019) partner repository with the dataset identifier PXD016388. Our application is hosted at https://shiny.imetalab.ca/pepFunk/ with source code and a read me file explaining package and version requirements available at https://github.com/northomics/pepFunk.

2.4 Datasets

2.4.1 Dataset 1: fecal microbiome treated with a histone deacetylace inhibitor

We adopted an *ex vivo* microbiome assay, termed rapid assay of individual's microbiome (RapidAIM) (Li *et al.*, 2019) to assess the direct effects of the histone deacetylace (HDAC) inhibitor suberoulanilide hydroxamic acid (SAHA) on a human microbiome. Briefly, in a RapidAIM assay, a human gut microbiome (fecal) sample was cultured for 24 h in anaerobic conditions in control conditions with dimethyl sulfoxide (DMSO) and treatment conditions with a low (0.125 mg/ml), or high concentration of SAHA (0.25 mg/ml). The human stool sampling protocol (Protocol # 20160585-01H) was approved by the Ottawa Health Science Network Research Ethics Board at the Ottawa Hospital. Proteins were digested with trypsin (Worthington Biochemical Corp., Lakewood, NJ). The digest was then desalted and analyzed using an Orbitrap Q-Exactive mass spectrometer as described previously (Zhang *et al.*, 2018a). Spectra search and peptide quantitation were completed using MetaLab v1.1.1 (Cheng *et al.*, 2017) and a database search of the IGC. The MS proteomics data have been deposited to the ProteomeXchange Consortium (Deutsch *et al.*, 2017) via the PRIDE (Perez-Riverol *et al.*, 2019) partner repository with the dataset identifier PXD016388. Peptide and protein group output files were used for the analyses. We removed peptides from the analysis if they were quantified in <50% of any condition. Of the total detected peptides, 79.5% were found in our core peptide database and 52.0% had at least one associated KEGG term (Supplementary Fig. S2A).

2.4.2 Dataset 2: fecal microbiome treated with metformin

A human fecal sample (microbiome) stored at −80°C was thawed quickly at 37°C and cultured using RapidAIM (Li *et al.*, 2019) for 24 h with 10 mM metformin (MTFM) or DMSO as the control. Control and treatment samples were cultured in five replicates. The human stool sampling protocol (Protocol # 20160585-01H) was approved by the Ottawa Health Science Network Research Ethics Board at the Ottawa Hospital. Cultured microbiome samples were subjected to protein extraction and tryptic digestion, and samples were analyzed using an Orbitrap Q-Exactive and a 90-min gradient as described previously (Li *et al.*, 2019). Three technical replicates were run on the Q-Exactive for a single sample (MTFM_3). Spectra search and peptide quantitation were completed using MetaLab v1.2.0 (Cheng *et al.*, 2017) using a database search of the IGC. The MS proteomics data have been deposited to the ProteomeXchange Consortium (Deutsch *et al.*, 2017) via the PRIDE (Perez-Riverol *et al.*, 2018) partner repository with the dataset identifier PXD016427. The median values of both peptide and protein-level intensities were used for the analyses for the three technical replicates (MTFM_3). We removed peptides from the analysis if they were quantified in <50% of either condition. Of the total detected peptides, 84.4% were found in our core peptide database and 62.4% had at least one associated KEGG term (Supplementary Fig. S2B).

2.4.3 Dataset 3: mucosal-luminal interface aspirates of pediatric patients with IBD

Raw sequence data from Zhang *et al.* (2018b), dataset identifier PXD007819, were downloaded from the ProteomeXchange Consortium (Deutsch *et al.*, 2017) via the PRIDE (Perez-Riverol *et al.*, 2018) partner repository. This study used mucosal-luminal interface (MLI) aspirates obtained from 71 pediatric (<18 years) patients with IBD. MLI aspirates were collected via colonoscopy at three intestinal locations: descending colon, ascending colon and the terminal ileum. For our study, we used a subset of the dataset to only include samples from the descending colon, totaling 62 samples [22 control, 22 diagnosed CD and 18 diagnosed ulcerative colitis (UC)]. Spectra search and peptide quantitation were completed using MetaLab and a database search of the IGC. We removed peptides from the analysis if they were quantified in <50% of any condition. Of the total detected peptides, 95.3% were found in our core peptide database and 53.0% had at least one associated KEGG term (Supplementary Fig. S2C).

3 Results

3.1 Sample group separation is possible using both peptide- and protein-level data

We first compared the ability principal component analysis (PCA) from intensity values at the protein- and peptide-levels distinguish treatment groups. To do so, we applied both protein and peptide-centric metaproteomic workflows to two fecal microbiome datasets: Dataset 1, a fecal microbiome treated with two concentrations of SAHA, Dataset 2, a fecal microbiome treated with MTFM and Dataset 3, MLI aspirates of patients with IBD. PCA using both peptide and protein group intensities is able to separate the high concentration of SAHA from the control DMSO treatment of Dataset 1 (Supplementary Fig. S1A and D). Clustering of treatments is similar using peptide and protein-level data analyses. In Dataset 2, PCA can also very clearly distinguish between a DMSO treated microbiome from one treated with MTFM using both peptide and protein group intensities (Supplementary Fig. S1B and E). The control DMSO treatment samples, however, cluster more tightly when using peptide intensities. Neither peptide- nor protein-level PCA was able to distinguish control patients from CD or UC (Supplementary Fig. S1C and F), which was expected and also identified by Zhang *et al.* (2018b).

3.2 KEGG functional enrichment of metaproteomic data

We compared protein- and peptide-centric workflows for KEGG pathway enrichment using a GSVA framework. Peptide spectra matches for both workflows were identified through a database search of the IGC peptide database using MetaLab (Cheng *et al.*, 2017). Our protein-level analysis considered protein groups that were identified by sequence similarity using MetaLab (Cheng *et al.*, 2017) and KEGG annotation for all proteins in each protein group were considered for functional enrichment.

We computed GSVA scores for all samples in each dataset. To test if GSVA scores followed the same trend in both workflows, we completed a correlation analysis of the median GSVA scores of pathways found to be significantly enriched at either a peptide- or protein-level analysis. We found linear agreement of GSVA scores between using protein- and peptide-level data sources with Pearson's correlation coefficients of 0.82, 0.85 and 0.69 for Datasets 1, 2 and 3, respectively (Fig. 2).

Using Dataset 1, we were able to complete GSVA on 74 and 91 protein and peptide sets, respectively. After protein and peptide set GSVA score ranking, a linear model in combination with an Empirical Bayes approach, was used to identify differentially

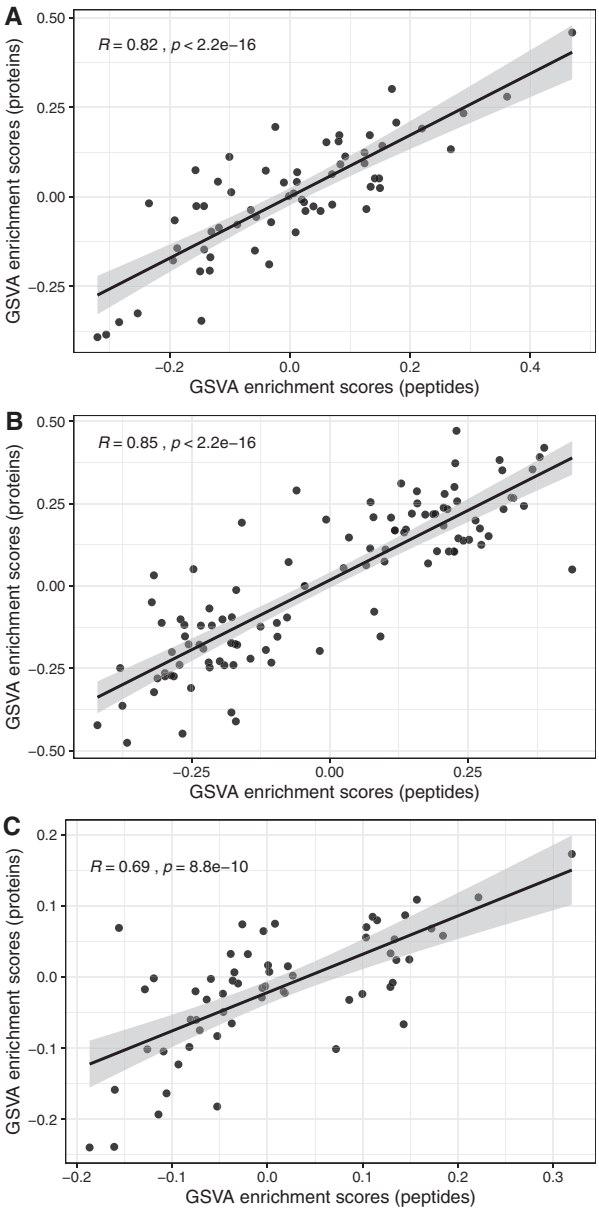

Fig. 2. Correlation of significant peptide-centric and protein-centric GSVA scores calculated in all three test datasets. Median GSVA scores at the condition level were used for the analysis. A linear regression line is plotted in yellow with a gray ribbon representing a 95% confidence interval. (**A**) Dataset 1, (**B**) Dataset 2 and (**C**) Dataset 3

enriched KEGG pathways in each of the treatment conditions (high and low SAHA) compared to a control. Using the protein-centric workflow on Dataset 1, we identified two consistent and significantly enriched KEGG pathways when comparing control DMSO treatment to both concentrations of SAHA [selenocompound metabolism (PATH: KO00450) and biosynthesis of ansamycins (PATH: KO01051)] (Fig. 3). By only comparing control DMSO conditions to high concentrations of SAHA treated microbiomes, we identified four additional enriched KEGG pathways.

The peptide-level analysis led us to identify 12 KEGG pathways as significantly enriched in samples treated with either low or high concentrations of SAHA (Fig. 3). Thiamin metabolism was the sole pathway that was only significantly enriched in samples treated with a low concentration of SAHA, while nine other KEGG pathways, such as glycerolipid metabolism and selenocompound metabolism, were significantly altered after treatment with high levels of SAHA. Five of the same pathways were identified as significant by both peptide- and protein-centric approaches; however, the peptide-centric approach was able to identify more significantly enriched KEGG pathways than the protein-centric method (Supplementary Fig. S3). Notably, significantly enriched KEGG pathways identified by the peptide-centric workflow were enriched in the same direction in the protein-centric workflow in both datasets (Fig. 3).

There was adequate detection of protein groups for GSVA analysis on 82 KEGG pathway protein sets. We identified 30 significant differentially enriched KEGG pathways in fecal microbiomes cultured with MTFM. Conversely, we were able to complete GSVA on 103 peptide sets using the peptide-centric approach, of which 47 were significantly enriched. Of the significantly enriched KEGG pathways, 24 were identified by both peptide- and protein-centric approaches (Supplementary Fig. S3C).

Because Dataset 3 was composed of gut microbiome data from 63 individual intra-condition variability between individuals was high (Fig. 5 and Supplementary Fig. S1C and F). Nonetheless, there was enough protein group quantitation for GSVA analysis on 101 and 98 pathway gene-sets using protein- and peptide-level data, respectively. In the peptide-level analysis, seven KEGG pathways were identified as significantly altered compared to control in CD microbiomes and five pathways in UC microbiomes (Fig. 5). Protein-level analysis identified slightly more or a similar number of altered KEGG pathways, with 11 differential pathways in CD microbiomes and 7 in UC samples (Fig. 5).

3.3 R shiny app

We created a web-based peptide-centric workflow made available as a companion tool to MetaLab (Cheng *et al.*, 2017) and iMetaLab (Liao *et al.*, 2018). Our application, pepFunk, accepts input files as MaxQuant peptide.txt files or user formatted files that include peptide sequence and intensity values. In addition, our application allows users to upload their own custom peptide-to-KEGG annotation file to extend the usability of this workflow to any proteomic experiment. The application performs the entire workflow and allows for the user to visualize data as a PCA biplot, a hierarchical dendrogram and two types of GSVA score heatplots. Users can also download analyzed data to create their own customized figures. The app is available at https://shiny.imetalab.ca/pepFunk with source code found at https://github.com/northomics/pepFunk. Dataset 1 has been provided as sample data.

4 Discussion

Functional analysis of metaproteomic data can be challenging. Database choice can have effects on the quality of results (Tanca *et al.*, 2016), redundant peptides can lead to ambiguously identified proteins (Ning *et al.*, 2016) and current methods of protein-level analysis can lead to a loss of information. Typically, protein-centric workflows are used and can be considered analogous to a transcriptomic workflow, where sequenced cDNA reads are mapped to genomic locations and analyses are completed on estimated transcript expression values. Recently, metatranscriptomics has moved toward functional annotation at cDNA read level which does not necessitate assembly or read mapping to genomic locations (Ugarte *et al.*, 2018). However, instead of enzymatic digestion as seen in metaproteomics, fragmentation of cDNA for metatranscriptomic sample preparation can be performed by physical methods, such as sonication (Marine *et al.*, 2011). The randomness of sonication typically leads to cDNA reads that map to unique locations in reference genomes, increasing the confidence of functional assignment.

In metaproteomics, the Protein Inference Problem, describing the challenges of peptide-to-protein assignment, can be even more challenging when considering the complexity of the microbiome. Because metaproteomic analyses can identify more proteins that share the same peptide sequences through the inclusion of multiple microbial strains and species, protein group-level analyses have been used to analyze proteins clustered into groups by sequence similarity. However, assigning peptides to protein groups leads to data loss

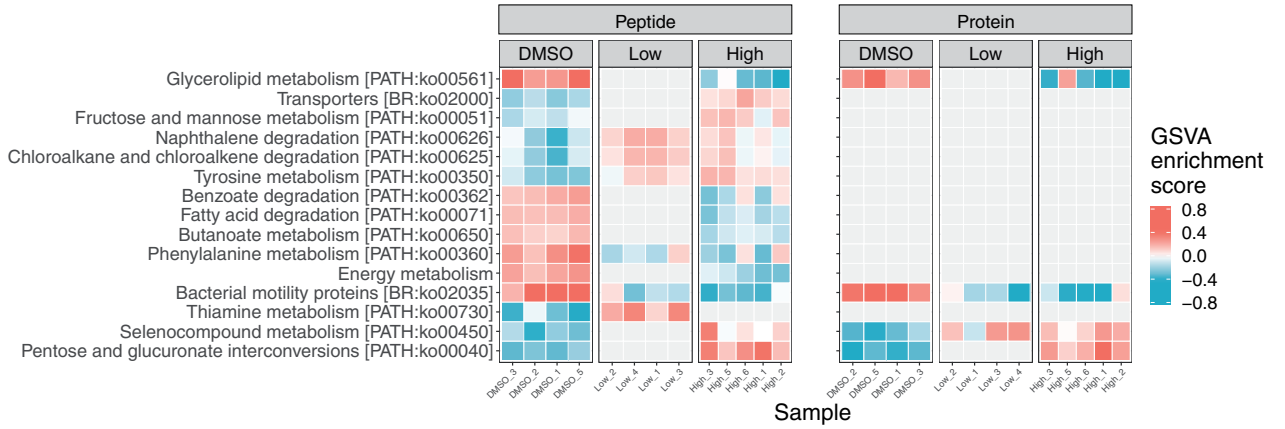

Fig. 3. Heatmap visualizing GSVA scores of Dataset 1. Peptide-centric workflow is presented on the left, and protein-centric on the right. A high score is visualized in coral and a low score in blue. A pathway is colored in gray if the pathway is enriched in one condition, but not the other (BH adjusted $P > 0.05$) (Color version of this figure is available at *Bioinformatics* online.)

where researchers can lose statistical power and potentially important functional information of their microbial community samples (e.g. Figs 3–5). To combat the issues that can arise from protein group pipelines, we have created a peptide-centric workflow. By analyzing metaproteomic data at the peptide level, we are able to identify similar enriched KEGG pathways as analysis at the protein group level. Furthermore, we can often identify more enriched KEGG pathways at the peptide level compared to the protein level because we retain more information (e.g. 12 versus 6 in Dataset 1 and 47 versus 31 in Dataset 2) (Figs 3 and 4 and Supplementary Fig. S3). Our peptide-centric workflow is unique as it uses a weighted intensity for functional assignment that is proportional to our confidence in annotated KEGG terms. In addition, our database is small and reduces computational resources required for a full functional database search.

To confirm the appropriateness of our approach, we looked at the biological relevance of the enriched KEGG pathways in our peptide-centric results. For Dataset 1, a microbiome treated with SAHA, an HDAC inhibitor, we looked at the cited functional roles of acetylation in bacteria. Acetylation is a reversible post-translational modification most well known for being essential to gene regulation. Acetylation can epigenetically alter expression by reducing the interaction between histones and DNA making DNA more accessible to transcriptional machinery (Sterner and Berger, 2000). Castaño-Cerezo *et al.* (2014) used *cobB* and *patZ* knockout mutants (a deacetylace and an acetyltransferase) in *Escherichia coli* to study how acetylation can affect the bacterial species. We expected to see similar functional results in our study to that by Castaño-Cerezo *et al.* (2014), particularly the results of the *cobB* knockout *E.coli*. Castaño-Cerezo *et al.* (2014) identified that 64% of acetylated proteins were associated to metabolism. Similarly, after SAHA treatment, both peptide- and protein-level analyses were able to identify alterations in expression to many pathways associated to metabolism, such as increases tyrosine (PATH: ko00350), selanocompound (PATH: KO00450) metabolism, and decreases in butanotate (PATH: KO00650), phenylalanine (PATH: KO00360), and benzoate (PATH: KO00362) metabolism (Fig. 3). Protein-level analysis was unable to identify the expression modulations in six of the seven altered KEGG pathways associated to metabolism. Castaño-Cerezo *et al.* (2014) also demonstrated that acetate metabolism itself is affected by acetylation through acetyl-CoA generation by acetyl-CoA synthase (ACS). For example, CobB was shown to preferentially deacetylate ACS, increasing its activity. The *cobB* mutant displayed reduced ACS activity, thus suggesting a reduction in acetyl-CoA generation. Using the peptide-centric analysis, we identified a reduction in both fatty acid degradation (PATH: KO00071) and butanoate metabolism (PATH: KO00362) pathways (Fig. 3), both of which result in acetyl-CoA. Reduced acetyl-CoA by HDAC inhibition may also be decreasing acetate metabolism. Neither of these pathways was identified by our protein-level analyses. However, our peptide-level analysis identified a reduction in cell motility when gut microbes were treated with SAHA, the opposite finding of Castaño-Cerezo *et al.* (2014).

MTFM, a drug, widely used in the treatment of type-II diabetes, has previously been shown to alter gut microbiome taxonomic composition and functionality (De La Cuesta-Zuluaga *et al.*, 2017; Li *et al.*, 2019; Ma *et al.*, 2018). In our study, both peptide- and protein-level identified MTFM-induced alterations to the same pathways as other studies. For example, we identified an increase in fatty acid biosynthesis (PATH: KO00061) (Li *et al.*, 2019) and tRNA biosynthesis (BR: KO03016) (Ma *et al.*, 2018). However, protein-level analysis could not identify other key pathways altered by MTFM treatment, such as a decrease in fructose and mannose metabolism (PATH: KO00051) (Li *et al.*, 2019). MTFM has also been shown to reduce folate metabolism (Cabreiro *et al.*, 2013), thus lower 'one carbon pool by folate' GSVA scores computed using both peptide and protein data of MTFM-treated samples are expected. However, peptide-level analysis identified a significant reduction in peptide intensity associated to glycolysis/gluconeogenesis in our MTFM-treated samples, the inverse of the findings by Li *et al.* (2019). This finding may be due to an abundance of proteolytic bacteria in this sample, or to the comparison between our human samples to the mouse samples from Li *et al.* (2019).

The gut microbiome is thought to play an important role in the biogenesis of both CD and UC. Interestingly, the gut microbiota of patients with IBD are known to be highly variable over time and can shift to a temporary 'healthy' state (Willing *et al.*, 2009), which may lead to the difficulty of distinguishing IBD from controls using both protein- and peptide-level analysis. However, both protein- and peptide-level analysis identified pathways that were up- or down-regulated in IBD compared to control individuals. Overall, both peptide- and protein-centric analyses revealed an average increase in oxidative phosphorylation (PATH: KO00190) in patients with CD and UC (Fig. 5). This corroborates with the study by Zhang *et al.* (2018b), which identified higher expression of proteins associated with DNA damage due to oxidative stress in the gut microbiota of IBD patients. This may also explain the down-regulation of genes associated with chromosome and associated proteins (BR: KO03036) identified in CD patients using peptide-level data. By using our peptide-centric approach, we also identified taurine and hypotaurine metabolism (PATH: KO00430), metabolites known to have antioxidant activity, as being significantly up-regulated in UC compared to control, which was not revealed by protein-level analysis. Similarly, using metabolomics, Kolho *et al.* (2017) reported that fecal taurine levels were significantly correlated with inflammation in pediatric patients with UC. Our peptide-level results, together with the above-mentioned metabolomic findings in pediatric UC patients, may suggest new mechanisms of host-microbiome interactions in response to the elevated oxidative stress in the gut lumen during the onset of UC.

Fig. 4. Heatmap visualizing GSVA scores of Dataset 2. Peptide-centric workflow is presented on the left, and protein-centric on the right. A high score is visualized in coral and a low score in blue. A pathway is colored in gray if the pathway is enriched in one condition, but not the other (BH adjusted $P > 0.05$) (Color version of this figure is available at *Bioinformatics* online.)

Currently, the vast majority of gut microbiome studies focus on using genomic sequencing to identify microorganisms. This type of meta'-omics' is useful at identifying the composition of a microbial community or its corresponding functional potential using a shotgun metagenomics approach (Halfvarson *et al.*, 2017). However, metagenomics cannot identify if genes are expressed and functionally active. Identifying the functionality of a microbiome sample is essential because multiple taxa can have redundant functions, and it is possible that a microbiome persists functionally even when taxa composition is altered (Blakeley-Ruiz *et al.*, 2019). As such, the emerging field of metaproteomics instead offers a functional snapshot of microbiome by identifying and quantifying translated proteins. Recently, multi-omic studies have shown that taxonomic variation, identified through metagenomics, may not always be associated with overall metaproteomic-identified functional changes in microbiome studies (Blakeley-Ruiz *et al.*, 2019; Mikan *et al.*, 2020). For example, Blakeley-Ruiz *et al.* (2019) identified compositional taxonomic changes in the microbiomes of patients with IBD yet persistent metabolic functionality both within and between patients. While it is accepted that the taxonomic composition and abundance in gut microbiomes are variable between individuals (Yatsunenko *et al.*, 2012), it is possible that redundancy in microbe functions result in an invariable metabolic landscape between individuals. Thus, if taxonomic changes do not always lead to functional shifts, metaproteomic studies should also consider taxa-independent functional analyses of samples, a focus of our peptide-centric workflow. Additionally, peptide-centric taxonomic analysis of metaproteomic data is already implemented by MetaLab (Cheng *et al.*, 2017), therefore, our peptide functional analysis completes the data analysis workflow and demonstrates the merit of working toward completely peptide-centric metaproteomic data analysis in gut microbiome studies.

GSVA, our functional analysis method of choice, can be used in a condition-independent manner in addition to comparing between control and treatment samples (Hänzelmann *et al.*, 2013). A condition-independent type of analysis is useful for researchers using functional pathways for exploratory analysis or for observing pathways that may be highly or lowly expressed in any given sample. Hänzelmann *et al.* (2013) also demonstrated that GSVA analysis has

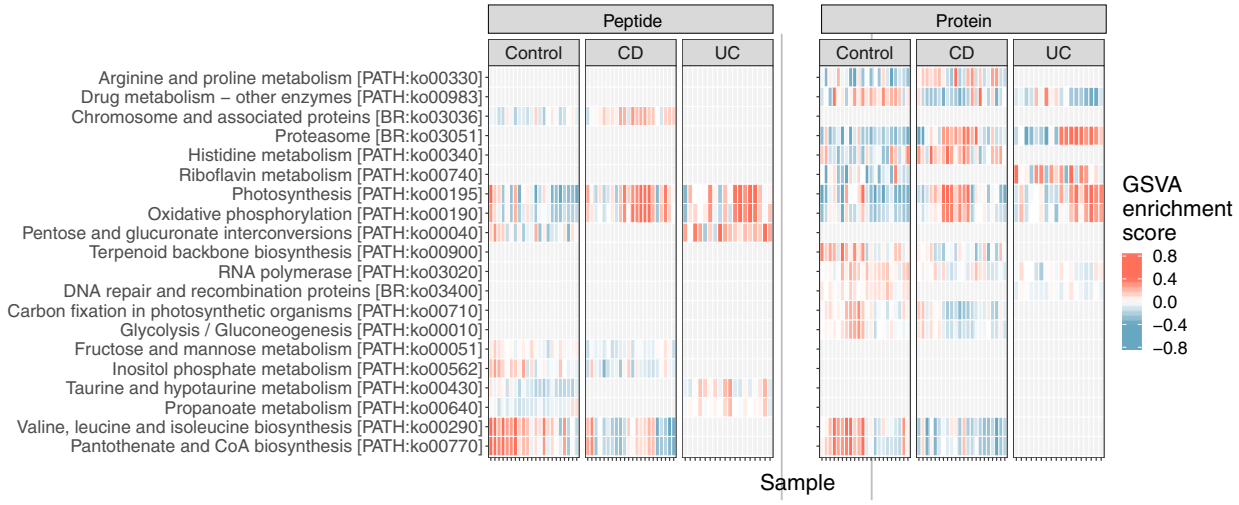

Fig. 5. Heatmap visualizing GSVA scores of Dataset 3. Peptide-centric workflow is presented on the left, and protein-centric on the right. A high score is visualized in coral and a low score in blue. A pathway is colored in gray if the pathway is enriched in one condition, but not the other (BH adjusted $P > 0.05$) (Color version of this figure is available at *Bioinformatics* online.)

a higher degree of sensitivity than other gene set enrichment techniques, such as ssGSEA and PLAGE, while simultaneously maintaining a low type-I error rate of ~0.05. While differential protein expression analysis can be used to study metaproteomic data (Hamann *et al.*, 2016), differential expression analysis can only identify differences of expression of individual proteins. Our implementation of GSVA gives sample- or experiment-wide interpretable results associated to KEGG pathways with a functional database specific to human gut microbiome studies. Currently, there also exists PSEA-Quant, which is a protein-centric gene set enrichment tool capable of performing both condition dependent and independent analyses, that uses protein set enrichment analysis, a method similar to our study (Lavallée-Adam *et al.*, 2014, 2015; Lavallée-Adam and Yates, 2016). However PSEA-Quant, uses protein-level data from proteomic studies, and is not specific to human gut microbiome experiments.

A limitation for peptide-centric approaches to functional enrichment analysis is the potential for functional over-representation of large proteins. We adapted our workflow from a classic GSVA analysis in attempt to reduce the negative impact of this limitation. Specifically, we used all possible KEGG annotations of each peptide and weighting peptide intensities accordingly. Because it is currently not always possible to know the parent-protein of each peptide, we believe our approach is appropriate for mitigating the possible challenges arising from peptide-level analysis. In addition, it is possible that our core database may be limiting the functional enrichment strategies we implemented in pepFunk. While many of the peptides identified by MS have potential KEGG terms that were not used in our analyses, these peptides represent a small portion of the total peptides identified in our analysis (Supplementary Fig. S2D–F). To address this concern, the R shiny app accepts a custom peptide-to-KEGG database allowing researchers to extend the microbiome-focused functionality of our core functional database to any type of metaproteomic experiment. The implementation of custom functional databases can now allow a user to use the most appropriate database version for their data. This includes the ability to use an algorithmically-made database if wanted. This type of database, however, is beyond the scope of this project as this tool was made with the gut microbiome metaproteomic community in mind.

As the field of metaproteomics grows, so does the need for accurate, fast and user-friendly tools for data analysis. Current protein focused functional enrichment workflows struggle with data loss stemming from assigning redundant peptides to proteins or protein groups. To combat this challenge, we created and implemented pepFunk, a peptide-centric functional enrichment workflow and accompanying R shiny tool for accurate and customizable data analysis. The current version includes a custom KEGG to human gut microbiome peptide functional database, but more experienced users can use their own annotated peptide database. As Ning *et al.* (2016) proposed, we have developed a workflow that directly analyses peptide intensities and is able to identify enriched KEGG pathways while maintaining the statistical validity of a protein-centric approach.

Acknowledgement

D.F. acknowledges a Distinguished Research Chair from the University of Ottawa.

Funding

This work was supported by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC)-CREATE TECHNOMISE program, the Government of Canada through Genome Canada and the Ontario Genomics Institute [OGI-156]; and the Province of Ontario. M.L.-A. holds an NSERC Discovery Grant. CMAS was funded by a stipend from the NSERC CREATE in Technologies for Microbiome Science and Engineering (TECHNOMISE) Program.

Conflict of Interest: D.F. co-founded Biotagenics and MedBiome, clinical microbiomics companies. The remaining authors declare no competing interests.

References

Arrieta,M.-C. *et al.*; the CHILD Study Investigators. (2015) Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Med.*, **7**, 307ra152.

Attali,D. (2017) *colourpicker: A Colour Picker Tool for Shiny and for Selecting Colours in Plots*. R package version 1.0. (19 March 2020, date last accessed).

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.*, **57**, 289–300.

Bjork,J. (2019) *LaCroixColoR: LaCroix Water Color Palettes*. https://github.com/johannesbjork/LaCroixColoR (19 March 2020, date last accessed).

Blakeley-Ruiz,J.A. *et al.* (2019) Metaproteomics reveals persistent and phylum-redundant metabolic functional stability in adult human gut microbiomes of Crohn's remission patients despite temporal variations in microbial taxa, genomes, and proteomes. *Microbiome*, **7**, 18.

Buchfink,B. *et al.* (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.

Cabreiro,F. *et al.* (2013) Metformin retards aging in *C. elegans* by altering microbial folate and methionine metabolism. *Cell*, **153**, 228–239.

Castaño-Cerezo,S. *et al.* (2014) Protein acetylation affects acetate metabolism, motility and acid stress response in *Escherichia coli*. *Mol. Syst. Biol.*, **10**, 762.

Chang,W. and Borges Ribeiro,B. (2018) *shinydashboard: Create Dashboards with 'Shiny'*. R package version 0.7.1. https://cran.r-project.org/package=shinydashboard. (19 March 2020, date last accessed).

Chang,W. *et al.* (2019) *shiny: Web Application Framework for R*. R package version 1.3.2. project.org/package=shiny. (10 November 2019, date last accessed).

Cheng,K. *et al.* (2017) MetaLab: an automated pipeline for metaproteomic data analysis. *Microbiome*, **5**, 157.

Cheng,K. *et al.* (2018) Separation and characterization of human microbiomes by metaproteomics. *Trends. Anal. Chem.*, **108**, 221–230.

Dash,S. *et al.* (2015) The gut microbiome and diet in psychiatry: focus on depression. *Curr. Opin. Psychiatry*, **28**, 1–6.

De La Cuesta-Zuluaga,J. *et al.* (2017) Metformin is associated with higher relative abundance of mucin-degrading akkermansia muciniphila and several short-chain fatty acid–producing microbiota in the gut. *Diabetes Care*, **40**, 54–62.

de Vries,A. and Ripley,B.D. (2016) *ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'*. R package version 0.1-20. https://cran.r-project.org/package=ggdendro. (19 March 2020, date last accessed).

Deutsch,E.W. *et al.* (2017) The proteomexchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.*, **45**, D1100–D1106.

Galili,T. (2015) dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*, **31**, 3718–3720.

Gaudet,P. and Dessimoz,C. (2016) Gene Ontology: pitfalls, biases and remedies. In: Dessimoz,C. and Skunca,N. (eds) *The Gene Ontology Handbook. Methods in Molecular Biology*. Vol. **1446**. Humana Press, New York, NY, pp. 189–205.

Gurdeep Singh,R. *et al.* (2019) Unipept 4.0: functional analysis of metaproteome data. *J. Proteome Res.*, **18**, 606–615.

Halfvarson,J. *et al.* (2017) Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.*, **2**, 17004.

Hamann,E. *et al.* (2016) Environmental Breviatea harbour mutualistic *Arcobacter* epibionts. *Nature*, **534**, 254–258.

Hänzelmann,S. *et al.* (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.

Hettich,R.L. *et al.* (2013) Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities. *Anal. Chem.*, **85**, 4203–4214.

Jangi,S. *et al.* (2016) Alterations of the human gut microbiome in multiple sclerosis. *Nat. Commun.*, **7**, 12015.

Kolho,K.-L. *et al.* (2017) Faecal and serum metabolomics in paediatric inflammatory bowel disease. *J. Crohns Colitis*, **11**, 321–334.

Lavallée-Adam,M. and Yates,J.R.,III (2016) Using PSEA-Quant for protein set enrichment analysis of quantitative mass spectrometry-based proteomics. *Curr. Protoc. Bioinformatics*, **53**, 13–28.

Lavallée-Adam,M. *et al.* (2014) PSEA-Quant: a protein set enrichment analysis on label-free and label-based protein quantification data. *J. Proteome Res.*, **13**, 5496–5509.

Lavallée-Adam,M. *et al.* (2015) From raw data to biological discoveries: a computational analysis pipeline for mass spectrometry-based proteomics. *J. Am. Soc. Mass Spectrom.*, **26**, 1820–1826.

Li,J. *et al.*; MetaHIT Consortium. (2014) An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.*, **32**, 834–841.

Li,L. *et al.* (2019) An in vitro model maintaining taxon-specific functional activities of the gut microbiome. *Nat. Commun.*, **10**, 4146.

Liao,B. *et al.* (2018) iMetaLab 1.0: a web platform for metaproteomics data analysis. *Bioinformatics*, **34**, 3954–3956.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Ma,W. *et al.* (2018) Metformin alters gut microbiota of healthy mice: implication for its potential role in gut microbiota homeostasis. *Front. Microbiol.*, **9**, 1336.

Marine,R. *et al.* (2011) Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Appl. Environ. Microbiol.*, **77**, 8071–8079.

Mikan,M.P. *et al.* (2020) Metaproteomics reveal that rapid perturbations in organic matter prioritize functional restructuring over taxonomy in western Arctic Ocean microbiomes. *ISME J.*, **14**, 39–52.

Moon,C. *et al.* (2018) Metaproteomics of colonic microbiota unveils discrete protein functions among colitic mice and control groups. *Proteomics*, **18**, 1700391.

Morgan,X.C. *et al.* (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.*, **13**, R79.

Muth,T. *et al.* (2015) The metaproteomeanalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. *J. Proteome Res.*, **14**, 1557–1565.

Nesvizhskii,A.I. and Aebersold,R. (2005) Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteom.*, **4**, 1419–1440.

Ning,Z. *et al.* (2016) Peptide-centric approaches provide an alternative perspective to re-examine quantitative proteomic data. *Anal. Chem.*, **88**, 1973–1978.

Owen,J. (2018) *rhandsontable: Interface to the 'Handsontable.js' Library*. R package version 0.3.7. https://cran.r-project.org/package=rhandsontable. (19 March 2020, date last accessed).

Perez-Riverol,Y. *et al.* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.*, **47**, D442–D450.

Perrier,V. *et al.* (2019) *shinyWidgets: Custom Inputs Widgets for Shiny*. R package version 0.4.9. https://cran.r-project.org/package=shinyWidgets. (10 November 2019, date last accessed).

Phipson,B. *et al.* (2016) Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Ann. Appl. Stat.*, **10**, 946–963.

R Core Team (2019) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Riffle,M. *et al.* (2018) MetaGOmics: a web-based tool for peptide-centric functional and taxonomic analysis of metaproteomics data. *Proteomes*, **6**, 2.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

Serang,O. and Noble,W. (2012) A review of statistical methods for protein identification using tandem mass spectrometry. *Stat. Interface*, **5**, 3–20.

Sievert,C. (2018) *plotly for R. https://cran.r-project.org/package=plotly*. (19 March 2020, date last accessed).

Sonnenburg,J.L. and Bäckhed,F. (2016) Diet–microbiota interactions as moderators of human metabolism. *Nature*, **535**, 56–64.

Starke,R. *et al.* (2019) Using proteins to study how microbes contribute to soil ecosystem services: the current state and future perspectives of soil metaproteomics. *J. Proteomics*, **198**, 50–58.

Sterner,D.E. and Berger,S.L. (2000) Acetylation of histones and transcription-related factors. *Microbiol. Mol. Biol. Rev.*, **64**, 435–459.

Suzek,B.E. *et al.*; the UniProt Consortium. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.

Tanca,A. *et al.* (2016) The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome*, **4**, 51.

Tang,W.W. *et al.* (2017) Gut microbiota in cardiovascular health and disease. *Circ. Res.*, **120**, 1183–1196.

Ugarte,A. *et al.* (2018) A multi-source domain annotation pipeline for quantitative metagenomic and metatranscriptomic functional profiling. *Microbiome*, **6**, 149.

Wickham,H. (2007) Reshaping data with the reshape package. *J. Stat. Softw.*, **21**, 1–20.

Wickham,H. (2011) The split-apply-combine strategy for data analysis. *J. Stat. Softw.*, **40**, 1–29.

Wickham,H. (2017) *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1. https://cran.r-project.org/package=tidyverse. (10 November 2019, date last accessed).

Willing,B. *et al.* (2009) Twin studies reveal specific imbalances in the mucosaassociated microbiota of patients with ileal Crohn's disease. *Inflamm. Bowel Dis.*, **15**, 653–660.

Xie,Y. *et al.* (2019) *DT: A Wrapper of the JavaScript Library 'DataTables'*. R package version 0.8. https://CRAN.R-project.org/package=DT. (10 November 2019, date last accessed).

Yatsunenko,T. *et al.* (2012) Human gut microbiome viewed across age and geography. *Nature*, **486**, 222–227.

Zhang,X. *et al.* (2018a) Assessing the impact of protein extraction methods for human gut metaproteomics. *J. Proteomics*, **180**, 120–127.

Zhang,X. *et al.* (2018b) Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nat. Commun.*, **9**, 2873.