Microbiome-wide characterization of post-translational modifications

Kai Cheng¹, Zhibin Ning¹, Xu Zhang¹, Janice Mayne¹ and Daniel Figeys¹,²*

*Correspondence: dfigeys@uottawa.ca

¹Department of Biochemistry, Microbiology and Immunology, Ottawa Institute of Systems Biology and Shanghai Institute of Materia Medica-University of Ottawa Joint Center in Systems and Personalized Pharmacology, Faculty of Medicine, University of Ottawa, Ontario, Canada

²Molecular Architecture of Life Program, Canadian Institute for Advanced Research, Toronto, Ontario, Canada.

Studying the structure and function of microbiomes is an emerging research field. Metaproteomic approaches focusing on the characterization of expressed proteins and post-translational modifications (PTM) provide deeper understanding of microbial communities. Previous research has highlighted the value of examining microbiome-wide protein expression in studying the roles of the microbiome in human diseases. Nevertheless, regulation of protein functions in complex microbiomes remains under-explored. This is mainly due to the lack of efficient bioinformatics tools to identify and quantify PTMs in the microbiome. Here we implemented an unrestricted open search strategy and developed an advanced multistage post-validation workflow for the identification of PTMs with high accuracy and sensitivity. This approach is applicable to proteomic and metaproteomic datasets. For metaproteomics, this method was combined into a complete data post-analysis workflow, including a quantitative module, to provide a novel and comprehensive data processing platform for metaproteomic analysis. This bioinformatics platform provides information about proteins, PTMs, taxa, functions, and pathways of microbial communities. The performance of the workflow was evaluated using conventional proteomics, metaproteomics from mouse and human gut microbiomes, and modification-specific enriched datasets.
The relationship between the microbiome and human health has been a very active subject of research in recent years\textsuperscript{1,2}. The study of host-microbiome relationships has been primarily focused on the composition of the microbial community and the relative abundance of microbes using DNA sequencing-based technologies, as well as functions of the microbiome using metagenomics and metatranscriptomics\textsuperscript{3}. Metaproteomics provides direct assessments of the microbiome functions by qualitative and quantitative investigating the expressed proteins products\textsuperscript{4,5-8}. Post-translational modifications (PTM), which are important for the dynamic regulations of the biological functions of the microbiome, can also be identified by mass spectrometry based metaproteomics\textsuperscript{9}. Yet, characterization of the PTM profile from microbial communities bring additional computational challenges in data interpretation. Conventional protein identification strategies allow a few pre-selected PTMs to be identified (variable modifications) but are not suitable to identify multiple modifications and unknown modifications. In contrast, open search strategies allow unbiased identification of potential PTMs and amino acid variants\textsuperscript{10}. Unfortunately, the application of open search strategies in metaproteomics has been limited. Open search strategies increase the search space when considering multiple possible mass differences for modified peptides. This is compounded for metaproteomics as protein sequence databases are already very large. Generally, an increase of several hundred-fold in processing time occurs during open searches. Fortunately, several recently developed software tools have significantly improved open search speeds\textsuperscript{11,12}, making open search feasible on a metaproteomic scale. Nevertheless, considering the complexity of microbial communities, ensuring the accuracy of peptide and PTM identifications is difficult. Several hundred potential PTMs can be found from one metaproteomic dataset. Calculation of the false discovery rate (FDR) at the peptide spectra match (PSM) level provides only an overall estimation of the confidence of identified peptides.

To address this challenge, we developed a multistage filtering strategy for the evaluation of open search results. A series of strict filtering criteria were applied for PTM, PSM, peptide and protein levels, which significantly improved the sensitivity and accuracy of the open search results. Since quantitative analysis is an indispensable step in proteomic data processing workflows, we incorporated a quantitative tool, FlashLFQ\textsuperscript{13} into our pipeline. By combining this tool with our previously developed taxonomy analysis and functional annotation modules\textsuperscript{14,15}, we built a complete workflow, which enables researchers to study the profile of PTMs of microbiota samples, as well as
relationships with taxa, functions and pathways. Moreover, traditional proteomic studies using single species are certainly supported by this strategy. This software is free for academic use and has a user-friendly Graphic User Interface (GUI). Both online and local versions of the tool are available at

https://imetalab.ca/.

RESULTS AND DISCUSSION

Tackling the problem of false discovery rate control with a multistage filtering strategy

The validation of peptide identification is a tricky problem in proteomics. The most prevalent strategy for validation of methods in proteomic software is the use of a target-decoy database search, which provides an estimate of the probability of random matches. Generally a FDR cutoff at 1% is accepted, however, it has been proposed that a lower FDR threshold (of about 0.1%) is required for open search results, which would ensure a comparable confidence level with the conventional closed search strategy\textsuperscript{16}. Moreover, only controlling the FDR at the PSM level is not sufficient for the investigation of PTMs.

Herein we developed a multistage filtering strategy for the confident identification of peptides and PTMs using open search (Figure 1). Briefly, first the MS/MS spectra were searched against a target-decoy database by open search. Then PSMs were binned according to their mass deviations (ΔMass) (Figure 2a). The ΔMass peaks were detected by Gaussian fitting analysis through the whole ΔMass range. The PSMs that did not belong to any ΔMass peaks were filtered out. Then for each Gaussian peak the \( R^2 \) was calculated and only peaks with \( R^2 \) above 0.9 were retained. The centroid of each peak was also determined during the curve fitting step, which was the experimental mass shift of the potential modification. Noting that the causes for the mass shift are various, including PTM, amino acid substitution, chemical derivative etc., they are all termed as “potential modifications”.

Usually in closed searches the precursor mass tolerance is set as 10 ppm. Therefore, in this step, the PSMs with ΔMass from the centroid of the corresponding Gaussian peak exceeding 10 ppm were discarded.

We first used a previously reported HEK293 data set\textsuperscript{10} for demonstration. After the PSM list was obtained and the Gaussian peaks along the ΔMass range were fitted, we investigated the score
distribution of the four types of PSMs: unmodified target PSMs (PSMs belonged to the Gaussian peak located at \( \Delta \text{Mass}=0 \), annotated as “peak 0”); target PSMs belonged to any other Gaussian peaks than the “peak 0” (“peak other”), for example, the 4 peaks annotated in Figure 2a; target PSMs outside of any Gaussian peaks (“no peak”); and all of the decoy PSMs (“decoy”) (Figure 2b). Overall the unmodified target PSMs were assigned with higher scores. The difference between “peak other” and “no peak” was major. However, the medians of “no peak” and “decoy” were quite close. This result showed that PSMs that didn’t belong to any \( \Delta \text{Mass} \) peak were more inclined to be random matches.

It was observed that PSMs that belonged to the fitted Gaussian peaks had higher scores than the PSMs outside of the Gaussian peaks (Figure 2c-m). For this dataset, 44.7% (463,673/1,038,110) PSMs didn’t belong to any Gaussian peaks. After this portion was removed, a FDR of about 1.1% (6,268/568,169) was obtained from the remaining PSMs. In this step 207 \( \Delta \text{Mass} \) peaks were kept, among which 160 peaks were identified as monoisotopic peptide peaks (Online Methods). As a result, aside from the peak with \( \Delta \text{Mass} = 0 \) representing unmodified peptides, 159 types of potential modifications were obtained. To further validate the result, we compared it with the closed search result. In the closed search 46,868 decoy PSMs were obtained in total, with 10,236 of the same spectra also identified as decoy matches in open search (Supplementary Fig. 1). This portion can represent random matches. We found these decoy matches were generally distributed evenly in the \( \Delta \text{Mass} \) range (about 10% were concentrated in \( \Delta \text{Mass} = 0 \) because unmodified peptides had higher priority). In the filtering step 84.2% (8,619/10,236) of them were removed because they didn’t belong to any \( \Delta \text{Mass} \) peak. This result proved that retaining only those PSMs belonging to high quality Gaussian peaks in the \( \Delta \text{Mass} \) range can effectively eliminate random matches. We also found that of 36,480 decoy PSMs in the closed search that matched to target sequences and that were assigned with higher scores, 46.5% (16,969/36,480) of them contained potential modifications which were determined in the previous step. Among the 16,969 PSMs assigned with potential modifications and higher scores, 6,601 PSMs were found with the potential modification \( \Delta \text{Mass}=1.0036 \), which should be the first isotopic peak. These results showed that the portion of modified peptides identified incorrectly in closed searches can be recovered by an open search strategy, improving the sensitivity of the analysis.

After the first filtering step the FDR was 1.1%, but it should be noted that if we calculated the modification-specific FDR, that is using the PSMs within a certain \( \Delta \text{Mass} \) Gaussian peak to calculate
the FDR of this group, the FDRs were below 1% for only 12 potential modifications (Supplementary Fig. 2). The FDR of the unmodified peptides which accounted for 80.2% (460,546/574,437) of the total amount of PSMs was only 0.28% (1,290/460,546). This meant that the FDR was higher for the modified portion. Only controlling the FDR at the level of PSM cannot assure the confidence of PSMs with certain modifications. Therefore, we added filtering at the level of modification. Firstly a classification score was assigned to each PSM based on a semi-supervised machine learning strategy (Online Methods, Supplementary Table 1). Then for each potential modification, an unpaired two-sample t-test was performed between target PSMs and decoy PSMs using the assigned classification score. If the two sets of score values showed significant difference (p-value < 0.01), this modification was retained. Following this step, 17 ∆Mass peaks were discarded. The remaining modifications were considered “true” with PSMs having significant differences between target and decoy matches.

The last filtering process was adopted at the protein and peptide levels. The protein inference algorithm was utilized at this step to calculate the scores of the proteins and peptides. Then the protein and peptide lists were filtered to ensure that, at both levels, the FDRs were less than 1%. For this dataset the final FDRs were only 0.08% (441/562,302) at the PSM level and 0.15% (179/120,037) at the peptide level, respectively (Supplementary Data 1). In total 142 possible modifications were detected, with 50 modifications matched with items in Unimod (http://www.unimod.org/) and 40 unknown modifications that can be constructed by combining two modifications from the 50 known modifications (Supplementary Table 2). We analyzed the compositions of 40 unknown modifications and found that 72% (29/40) of them contained at least one modification listed below: -156.1019, 156.1014, 128.0951 and -128.0958 Da, which corresponded to addition/loss of Arg or Lys. Loss of the terminal amino acid by in-source fragmentation was common in ESI-MS. Addition of Arg or Lys could be caused by additional missed cleavage sites (consecutive Arg or Lys in the peptide C-term). In addition, 25% (10/40) of the combinations contained -71.0375 Da (Gln->Gly), -71.0743 Da (Lys->Gly) or -57.0221 Da (Asn->Gly/Gln->Ala), which were related to amino acid substitutions. Therefore, from the combination of information we can determine that potential modifications without Unimod matched items, can be explained by known modifications to a large extent.

In summary, we applied this method to a large scale proteomic dataset from the HEK293 cell line. The results illustrated that high confident identifications at all of the four levels (PSMs, peptides,
proteins and potential modifications) were obtained. Compared with conventional database search methods, the sensitivity was significantly improved by our strategy (Supplementary Fig. 3a). Using this strategy 562,302 PSMs were identified with PSM level FDR<0.08%, and peptide and protein level FDR<1%. In a closed search in which even the peptide and protein level FDR were not considered, 434,526 PSMs were obtained with PSM level FDR<1%. If we continued to restrict the FDR at the peptide and protein levels to 1%, 385,990 PSMs remain (Supplementary Fig. 3b). The significant improvement of sensitivity using our strategy was not only because of the identification of modified peptides, which cannot be identified in the conventional method; in fact, the 459,367 unmodified PSMs obtained in open search were more than the number obtained in closed search. With proper filtering, the increased search space did not sacrifice the identification sensitivity; on the contrary, it improved it. The first reason was 36.2% (16,969/46,868) decoy PSMs were matched to more credible targets. This portion of loss was recovered in open search, at the same time the additional expense for the elimination of these random matches in FDR control was avoided. The second reason was that random matches were basically evenly distributed over a wider range, which made the ΔMass a significant feature to distinguish random matches effectively. Based on these reasons we developed the post-analysis workflow for the analysis of open search datasets, which showed greatly improved accuracy and sensitivity of identifications.

Identification of a group of phosphorylation related modifications from enriched samples

To further benchmark the post-validation strategy, we re-analyzed E. coli phospho-enriched samples (Dataset ID: Ecoli_phos) to measure the proportion of PSMs that could be identified as phosphopeptides. Authenticating our strategy, phosphorylation was identified as the most abundant modification from the phospho-enriched E. coli sample. Even more substantiating, the 2\textsuperscript{nd}-4\textsuperscript{th} potential modifications were phosphate-ribosylation (ΔMass=212.0091), PhosphoHex (ΔMass=242.0195) and 3-phosphoglyceryl (ΔMass=167.983), respectively, which were all phosphorylation related modifications. Actually, by manual interpretation of the spectra from the top 30 abundant mass differences, we found that neutral loss of phosphorylation was observed from 14 types of potential modifications (Supplementary Data 2, Supplementary Material 1). Aside from phosphorylation, pyrophosphorylation (ΔMass=159.9328), and the abundant phospho-related
modifications mentioned above, other known phospho-related modifications also include phosphogluconoylation (ΔMass=258.0149) and pyridoxal phosphate (ΔMass=229.0134). No Unimod records were found for possible modifications with experimental ΔMass 151.9878. We inferred the composition as H(5)C(3)O(5)P because according to the ΔMass it is most likely deducing one oxygen from 3-phosphoglyceryl. We also inferred the composition of ΔMass 181.9985 as H(7)C(4)O(6)P(1), because the mass difference between this and 3-phosphoglyceryl is 14.0154 Da which could be a methylene group. The possible structures of the identified phospho-related modifications were illustrated (Figure 3a).

Phosphate-ribosylation is a tag of ADP-Ribosylation, which is a versatile PTM involved in various cellular processes\(^\text{20, 21}\). It has been reported that ADP-Ribosylation sites could be detected in phosphoproteomics data by setting it as the corresponding variable modification during database searches\(^\text{22}\). In this dataset phosphate-ribosylation were identified as the 2\(^{nd}\) most frequent PSM count. In theory, the ADP-ribose group can be attached on multiple amino acids, such as Asp, Glu, Lys, Arg and Cys. The specificity sites for phosphate-ribosylation recorded in Unimod are Asp, Glu, and Arg. However, in this dataset, the modification sites were observed as Lys and protein N-term. In total, 86.3% (864/1001) of phosphate-ribosylation modified peptides contain a missed cleavage site Lys. Additionally, 7.4% (74/1001) of peptides start with the first amino acid of the proteins, and 8.9% (89/1001) of peptides start with the second amino acid of the proteins, for a total of 16.3% (163/1029) peptides containing protein N-term (Supplementary Material 1). Through the functional enrichment analysis of the proteins with phosphate-ribosylation sites, 61 biological processes with FDR below 0.05 were obtained (Figure 3b, Supplementary Table 3, 4). The colored three biological processes were all branches of nucleoside phosphate metabolic process, which suggested the important role of ADP-ribosylation in this process. The 3\(^{rd}\) and 4\(^{th}\) potential modifications were matched with PhosphoHex and 3-phosphoglyceryl, respectively. As above, most phosphoHex and 3-phosphoglyceryl modified peptides contained a missed cleavage site (Lys) or protein N-term. The manually annotated spectra showed that the modifications occurred on the Lys at the missed cleavage site or protein N-term (Supplementary material 1). These results clearly showed the advantage of open search for the identification of coexisting PTMs from enriched samples. For example, Lys and protein N-term were not specified sites of phosphate-ribosylation and phosphoHex in Unimod, so these sites would be missed when setting these as variable modifications in closed
searches. However, in open searches, novel modifications or known modifications on novel sites can be readily identified. The proteome-wide investigation of phosphate-ribosylation sites and their abundance is important for the further understanding of ADP-ribosylation functions.

Encouraged by this success we investigated phosphorylation enriched mammalian samples from mouse kidney (Dataset ID: Mouse_phos). Surprisingly, only two of the top 30 ∆Masses were matched with known modifications in Unimod, namely phosphorylation and pyrophosphorylation (Supplementary Data 3). It’s easy to find ∆Masses 239.8978 and 319.8637 that represented multiple phosphorylation sites occurring on one peptide. By manually interpretation of the spectra we found that all of the unknown ∆Masses were combinations of multiple modifications, and in most cases, were combinations of phosphorylation and others (Supplementary material 2). In fact, among the top 30 abundant ∆Masses, 26 contained at least one phosphate group, which strongly confirmed the confidence of this result. Four of them were combinations of phosphorylation and cations (protons replaced by Na⁺ or K⁺). This type was not observed for Ecoli_phos (but common in mouse_phos). While 3-phosphoglyceryl and phosphate-ribosylation were found with few PSMs counts and phosphoHex was not observed. Apart from this, peptides from mouse_phos tended to contain multiple modifications, which was a significant difference between these two datasets.

We also studied two human cell lines, HeLa (Dataset ID: Hela_phos, Supplementary Data 4, Supplementary material 3) and U-87 glioblastoma cells (Dataset ID: U87_phos, Supplementary Data 5, Supplementary material 4). Similar to the mouse_phos dataset, multiple phosphorylation and combinations of phosphorylation and other modifications were prevalent. But complex phospho-related modifications as these existed in Ecoli_phos data were very rarely observed. Our results clearly showed the different features of phosphopeptides identified from bacteria and mammal cells. In Ecoli_phos the majority of peptides have only one modification, including some complex modification forms and unknown modifications. From the mammalian datasets, multiple phosphorylation sites on one peptide were common rather than phospho-related complex modifications. At the same time, peptides modified by cations were often observed from mammalian datasets, but rare in Ecoli_phos datasets. Using our strategy co-enriched modifications can be identified in one step which is of great significance for the analysis of interactions of different modifications. Collectively, these results showed significant benefits of the application of open searches for enriched sample analyses.
Identifying the modification profile from microbiome samples

We also applied our workflow to analyze a microbiome dataset which was constructed by assembling 32 species/strains of Archaea, Bacteria, Eukaryotes and Bacteriophages\textsuperscript{26} (Supplementary Note, Supplementary Data 6). Given the advantage of known microbial composition in this dataset, we also used this dataset to benchmark the quantitative accuracy of our method, to investigate if the quantitation will be affected by the additionally identified peptides from open search. In total, we identified 1,048,759 PSMs, which was 16\% more than from conventional closed searches (Supplementary Fig. 4). Through quantitative analysis of the mock microbiota community, satisfactory quantitative accuracy was obtained at the species-level, which also confirmed the confidence of the identifications of modified peptides (Supplementary Fig. 5, Supplementary Note).

To demonstrate the capability of open searches for microbiome sample analyses we applied the developed strategy to analyze a mouse gut microbiome dataset (Dataset ID: Mouse_gut\textsuperscript{27}). In total 918,132 PSMs were identified with an estimated FDR of 0.17\%. The PSM count increased significantly by 28.3\%, 202,682 additional PSMs were identified when compared with our previous work \textsuperscript{14}. At the peptide level the FDR is 0.33\% and 127,218 unique sequences (37,544 peptides contained at least one modified form) were obtained (Supplementary Data 7). In closed search 115,832 peptide identifications were obtained and so the open search strategy resulted in a 9.8\% increase in peptide identifications (Figure 4a). For the commonly identified portion, most of the peptides were only identified as unmodified in the open search, but there were about 20\% peptides that contained at least one modified form. For the additional identified portion by open search, about 40\% of peptides were found only in modified forms and over 50\% were unmodified. This result showed that, not only was the missing information of modified peptides retrieved from the open search, but also, that the improved sensitivity of the workflow enabled the additional detection of unmodified peptides. The increase in the number of confidently recognized species (with at least three unique peptides) was also realized. In this work 172 species were identified containing equal or more than three unique peptides. The number was 143 from the previous work, with a major common portion of 134 species shared by the two results (Figure 4b). The 29 species identified only in this workflow were illustrated and their distribution at the phylum level was shown in Figure 4b.

Most of the newly found species were from phyla Firmicutes and Bacteroidetes. Actually, 27 species
were also identified using the conventional method but without enough unique peptide counts, which also showed the confidence of the species specifically determined using the new strategy. From this result we can conclude that through our strategy, comparable accuracy and optimized sensitivity can be obtained by open searches when compared with traditional closed searches for the analysis of real microbiota samples.

In this experiment the mouse gut microbiome were collected from both high-fat diet (HFD) and low-fat diet (LFD) fed mice for the interpretation of diet-microbiota-host interactions. In this work, based on the quantitative information of the potential modifications, we investigated the response of the determined modifications to diet in mouse microbiomes. The abundances of various modifications were calculated in each identified taxon node. Both the absolute and relative values were considered, that is, for each taxon the summation of the intensity of peptides with specific modifications were calculated as the absolute abundance of the modification in this taxon, and the ratio between the absolute value and the total abundance of this taxon were calculated as the relative abundance. Firstly, we tested if there were modifications that showed significant differences between HFD and LFD. We found that the modification $\Delta$Mass=71.0368 (Acrylamide adduct/Gly->Gln substitution) showed significant increases in the HFD samples using both absolute and relative values, and the modification $\Delta$Mass=52.1229 (replacement of 3 protons by iron) decreased significantly in relative value (Figure 5a, Supplementary material 5). This result confirmed the ability of our strategy for the identification of quantitative changes of PTMs from real gut microbiome samples.

Since there were two possibilities for the modification $\Delta$Mass=71.0368, we investigated the information of localization sites to determine which was validate. Acrylamide adducts may occur on Cys, Lys and any amino acid on peptide N-term whereas Gly->Gln substitution only exists on Gly. Considering all the possible sites for localization analysis, we found Lys the most frequent modification site for $\Delta$Mass=71.0368 (Supplementary Fig.6), which suggested that the most probable modification was an acrylamide adduct. The relative abundance of modified peptides in each taxonomy in both HFD and LFD samples were illustrated respectively (Figure 5b). The compositions and relative abundances were almost constant in these two conditions. The lineage Bacteria->Firmicutes->Bacilli->Lactobacillales->Streptococcaceae->Lactococcus->Lactococcus lactis was the main source for acrylamide adduct modifications and contributed to the high abundance in HFD samples directly. At the protein level the difference was derived from the single site K264 from
protein glyceraldehyde-3-phosphate dehydrogenase (GAPDH) in *Lactococcus lactis*. GAPDH is involved in the glucose metabolic process and related to the function NAD binding, NADP binding and GAPDH (NAD+) (phosphorylating) activity. Acrylamide adducts have been deemed as potentially toxic compounds in food\(^\text{28}\). The degradation of acrylamide by bacteria was widely studied\(^\text{29, 30}\) and *Bacillus cereus* has been reported to efficiently degrade acrylamide\(^\text{31}\). Molecular mechanisms of the addition of acrylamide adducts on GAPDH was studied by incubating purified human GAPDH with acrylamide and the kinetics of Cys acrylamide adduct formation was characterized, but the modification on Lys was not described\(^\text{32}\). This work presented the novel discovery of Lys acrylamide adducts on GAPDH from *Lactococcus lactis*, and the potential relationship with HFDs.

For the modification replacement of three protons by iron (cation:Fe[III] for short), the most abundant species were *Bacteroides vulgatus* and *Parabacteroides goldsteinii* from the same order *Bacteroidales* (Figure 5c). Significant change was only observed in species *Parabacteroides goldsteinii*. The involved proteins included TonB-dependent receptor, RagB/SusD family nutrient uptake outer membrane protein, SusC/RagA family TonB-linked outer membrane protein and hypothetical proteins. The first three were all membrane proteins and TonB-dependent receptor was a known iron transporter\(^\text{33}\). The identification of cation:Fe[III] modification on these proteins could help researchers further their understanding of the transport mechanism of bacteria.

**Identification of multiple types of glycosylation from human gut microbiome dataset**

We then applied this workflow for the analysis of a human gut microbiome dataset from a study about type 1 diabetes (Dataset ID: Human_gut)\(^\text{34}\). Two subject groups of data from healthy control (CO, n=22) and new-onset patients (NO, n=33) were analyzed. From this dataset we identified 829,458 PSMs at FDR about 0.16% (Supplementary Data 8). Only 676,282 PSMs were obtained at the FDR about 4.1% by the conventional method (Supplementary Fig. 7). The identification rate was increased from 35.8% to 44.0%, and while a much lower FDR was obtained. For each sample, 10,305 ± 2,205 (mean ± standard deviation) unique peptide sequences and 4,763 ± 1,027 proteins were quantified. In total 43,475 non-human microbiome proteins and 629 human proteins were identified.

We quantified 611 species, among which 240 species were identified with more than three unique peptides. Annotations with KEGG database allowed us to identify 459 ± 69 microbial and 100 ± 27
human KEGG orthologies (KOs) per sample.

Here we identified 288 types of potential modifications and 102 were successfully matched to records in Unimod. Those known modifications covered 82% (78,931/97,102) of the modified PSMs. In general, human proteins had more PTMs than microbial proteins (Supplementary Fig. 8). PSMs from human proteins accounting for 17.7% (146,734/829,458) of the total number but contributed to 41.8% (40,608/97,102) of the PSMs with PTMs. The dynamic range of various modifications was high. The top 10 modifications covered 56.6% of the total number of modified PSMs identified in this study (Supplementary Fig. 9).

Currently the glycosylation in the microbial community is an under-studied area. For a better understanding of the gut microbiome-host interactions the characterization of glycosylation is absolutely necessary. In this study, five types of glycosylation including Hex (hexose, ∆Mass = 162.053), HexNAc (N-Acetylhexosamine, ∆Mass = 203.079), HexNAc1dHex1 (∆Mass = 349.137), HexNAc2 (∆Mass = 406.159) and bacillosamine (2,4-diacetamido-2,4,6-trideoxyglucopyranose, ∆Mass = 228.1110) were identified from this dataset (Figure 6a). An important feature of the glycopeptide spectra were that the diagnostic ions could be observed in low mass-to-charge ratios (m/z), which were called glycan oxonium ions35. In our results, we detected at least one glycan oxonium ion whose intensity exceeded 50% of the peaks in the corresponding MS2 spectra, in 92.0% (5,424/5,895), 94.6% (5,750/6,078), 93.4% (157/168) and 69.5% (91/131) of the HexNAc, HexNAc1dHex1, HexNAc2, and bacillosamine modified PSMs, respectively (Supplementary Table 5). Manual verifications were also adopted for part of the glycopeptide spectra, and generally successive y-ions were observed in these spectra (Supplementary material 6). These results ensured the reliability of the identification of glycopeptides. Interestingly, we found the distributions of these glycosylations between human and microbial proteins differed. Two types of glycosylation with apparently higher abundances, HexNAc1dHex1 and HexNAc were only observed on human proteins. For Hex and HexNAc2, there were 38.1% (45/118) and 28.6% (48/168) of the PSMs that were identified from microbial proteins, respectively. Bacillosamine was a rare monosaccharide which was found in bacterial glycoproteins36. And in our results 93.1% (122/131) of the bacillosamine modified PSMs were from bacteria proteins.

Glycosylation plays a critical role in the regulation of protein functions. The alterations of glycosylation could affect the immune system of mammals37. The relationship between abnormal...
glycosylation and the pathogenesis of diabetes has also been reported\textsuperscript{38}. In this dataset, significant decreases in the abundance of HexNAc modified peptides were detected in NO patients compared to CO subjects (Figure 6b). Functional enrichment analysis of HexNAc modified proteins was performed (Figure 6c, Supplementary Table 6). Neutrophil related immune response was the most significantly enriched biological processes (Supplementary Fig. 10). Among the enriched Gene Ontology (GO) terms, nine of them were significantly changed (Supplementary Fig. 11). Five of them (transport, response to stress, inflammatory response, response to stimulus and localization) were also downregulated in NO samples which were consistent with the general trend. But four catabolic processes (catabolic process, cellular catabolic process, organonitrogen compound catabolic process and organic substance catabolic process) related functions were upregulated.

Besides high abundant human protein glycosylation, we also identified three types of glycosylation from microbial proteins. For hexose, 21 glycopeptides were identified, only one peptide was from Bacteroidetes and 16 peptides were from Firmicutes (Supplementary Fig. 12, Supplementary table 7). In this dataset, 86.7\% (39/45) of the hexose modified PSMs were from Clostridiales. From the functional annotation, we found that the most frequently observed function related to hexose was ABC-type glycerol-3-phosphate transport system, periplasmic component (Figure 6d). For HexNAc2, one peptide EAATTTTTTTTTDVADKK was identified from eight CO and seven NO samples, respectively, which was the most widely identified glycopeptide from bacteria in Human_gut (Supplementary Fig. 13, Supplementary table 7). The lowest common ancestor (LCA) of this peptide was Clostridia (order level). The corresponding protein was ABC-type sugar transport system, periplasmic component, contains N-terminal xre family HTH domain, which was also an ATP-binding cassette (ABC) transporter. It could be seen that the ABC transporter families were the most frequently identified glycoproteins from bacteria in Human_gut. We also identified bacteria specific monosaccharide Bacillosamine from 48 glycopeptides. Bacillosamine was described as an “Asn-linked glycan from Gram-negative Bacterium” in Unimod. However, we noticed that only 45.8\% (22/48) glycopeptides contained Asn, which suggested the existence of other types such as O-linked glycopeptides in these samples. Bacillosamine was involved in various functions in this dataset (Supplementary Table 7). We identified the LCA for 21 glycopeptides (Supplementary Fig. 14). Three peptides determined in superkingdom level as bacteria were from the protein Elongation factor Tu (EF-Tu), which was a highly conserved proteins in prokaryotes. EF-Tu was a type of GTPase, which
was also an important antibiotic target. For the remainder of the glycopeptides, seven were from Bacteroidetes and eight were from Firmicutes. The main COG categories were carbohydrate transport and metabolism (G) and energy production and conversion (C). The characterization of bacteria glycosylation will greatly improve our understanding of gut microbiome-host interactions.

Conclusion

Modifications of peptides bring mass variations which can increase the number of unidentified spectra in traditional database searches. The open search strategy is a method capable of identifying modifications in an unrestricted manner, however, the validation of the results remains challenging, especially for the studies of microbial communities. To address this, we developed a multistage filtering strategy, which could ensure the confidence of the results at various levels: PSMs, peptides, proteins and modifications. In this method, rigorous filters are utilized to remove potential modifications with less confidence. The analysis of the benchmark datasets (Homo_HEK293, Mock_micro, Mouse_gut and Human_gut) suggests that high accuracy is achieved without the sacrifice of sensitivity. It is worth noting that the identification rates are greatly improved in the analysis of complex mouse/human gut microbiota samples using open search strategy. Combined with the previously developed metaproteomics data analysis tools by our group, a novel bioinformatics workflow is provided for the comprehensive analysis of PTM profiles from microbial communities. Rich and quantitative information can be obtained easily and with high accuracy, which is the major reason we recommend this strategy to be a standard workflow for proteomics/metaproteomics studies.
Figures

Figure 1.
Workflow of the post-analysis processes of open search datasets. After the open search result was obtained, in steps 2-5 a multi-stage filtering strategy was adopted for the results validation. Then MS1 intensity-based label free quantification was performed. The final step was taxonomy analysis and functional annotation for metaproteomics studies. This workflow is also suited for the analysis of conventional proteomics datasets.

Figure 2.
(a) The density of ΔMass distribution of the identified PSMs. The zoomed-in part shows four ΔMass peaks which were detected by Gaussian fitting along the ΔMass range. (b-m) Score distributions of PSMs identified by open search. The score was calculated as -Log_{10}(Expect) and higher scores represented better matches. (b) Peak 0: the scores of the PSMs in the fitted Gaussian peak on the position ΔMass=0, which consisted of PSMs without modifications; Peak other: the PSMs in other fitted Gaussian peaks than Peak 0, which consisted of modified PSMs; No peak: the scores of the PSMs with modifications but do not belong to any ΔMass peaks; Decoy: the scores of all the decoy PSMs. (c) first box: the scores of the PSMs with ΔMass from -0.07 to -0.02. -0.07 was the right edge of a Gaussian peak and -0.02 was the left edge of the next Gaussian peak, therefore the PSMs in this range did not belong to any Gaussian peak; second box: the scores of the PSMs with ΔMass from -0.02 to 0.02. -0.02 and 0.02 were the left and right edges of a fitted Gaussian peak respectively, which consisted of unmodified PSMs; third box: the scores of the PSMs with ΔMass from 0.02 to 0.09. 0.09 was the left edge of the next Gaussian peak, so the PSMs in this range also do not belong to any Gaussian peak; fourth box: the scores of the decoy PSMs with ΔMass from -0.07 to 0.09. (d-m) examples of the scores of the PSMs with other potential modifications. Similarly, the first box represented the scores of the PSMs between this Gaussian peak and the previous Gaussian peak; the second box represented the scores of the PSMs belong to this Gaussian peak; the third box represented the scores of the PSMs between this Gaussian peak and the next Gaussian peak; the fourth box represented the scores of all the decoy PSMs in this range.

Figure 3.
Phospho-related modifications identified from phosphorylation enriched E.coli samples. (a) Possible structures and the corresponding PSM counts of phospho-related modifications; (b) network and enriched biological process of proteins with phosphate-ribosylation modifications.

Figure 4.
The comparison of the identified (a) peptides and (b) species by closed search and open search from
mouse gut microbiota samples.

**Figure 5.**

(a) The absolute and relative abundances of potential modifications $\Delta$Mass=71.0368 and $\Delta$Mass=52.1229 identified in mouse gut microbiome. (b-c) The relative abundance of each taxon determined by (b) acrylamide adduct and (c) cation:Fe[III] modified peptides in HFD and LFD samples.

**Figure 6.**

(a) Identified potential modifications and their distribution between human and microbial proteins. Purple dots: modifications matched in Unimod; Grey dots: unknown modifications. The annotated modifications fulfilled one of the following two conditions: 1) with above 1,000 PSM counts; 2) with above 100 PSM counts and more than 90% were from microbial proteins. The possible structures of five glycosylation were illustrated. (b) HexNAc modified glycopeptides showed significant difference between CO and NO samples. (c) The enriched biological process of HexNAc modified human glycoproteins. (d-e) The functions of the microbial proteins modified by (d) Hex and (e) HexNAc. CO: control subjects; NO: new-onset patients.
Methods

Workflow

This strategy has been incorporated into our previously developed metaproteomics data analysis software MetaLab (version 2.0.0). Only mass spectrometer raw files and searched protein databases were required as the input. The input raw files were converted to mzXML and MGF format for the post analysis. The converted mzXML files were subject to MSFragger (version 20171106) for the peptide identification. Carboxyamidomethylation on C was set as fixed modification. Although an open search can identify modifications unrestrictedly, considering that oxidation on Met and acetylation on protein N-term commonly exist, these two were still set as variable modification. Oxidation and acetylation that occur on other amino acids would be identified by open search. The precursor mass tolerance was set as 500 Da and the fragment mass tolerance was set as 10 ppm. A target-decoy database was utilized for the FDR estimation.

After the identification results were obtained, all the PSMs were binned according to the determined ∆Mass. Through the ∆Mass range, the Gaussian peaks with R² above 0.9 were detected (users can change this parameter in our software). PSMs don’t belong to any ∆Mass Gaussian peak were discarded. The centroids of the fitted Gaussian peaks were determined. For each PSM, the mass of their potential modification was set as the mass of the calculated centroid, instead of the mass deviation determined directly from the open search result. If the relative mass difference between the original mass deviation and the calculated centroid was above 10 ppm, this PSM was also discarded. After the mass of the modification of each PSM was determined, deisotoping was performed. The theory mass difference between two isotope peaks was set as 1.00286864 Da.

When we found two Gaussian peaks, Peak_i (∆Mass=M_i) and Peak_j (∆Mass=M_j), which fulfilled the conditions that |M_j-M_i-1.00286864|≤0.01, and at the same time Peak_i had more PSMs than Peak_j, Peak_j was deemed as an isotope peak of Peak_i. The mass of modifications of PSMs in Peak_j was set as M_i in this step.

Then a semi-supervised machine learning strategy was performed (Supplementary Table 1). The 10% top scored PSMs were randomly distributed into five datasets, and all the decoy PSMs were
appended to each of the datasets for the construction of training datasets. Five naive Bayes classifiers were built based on the training datasets and used to calculate the probability for each PSM to be a target/decoy one. A classification score was assigned for each PSM, which was calculated by the average of the probabilities to be a target match determined by the five classifiers. Then the PSMs were sorted by the newly assigned classification score, and the top 10% were selected for another loop of the classification. After five iterations, the average of the probabilities were determined as the final classification score.

The next step was filtering PSMs at the modification level. Unpaired t-tests were performed between the classification scores of target and decoy PSMs. We also created a universal decoy sample which consisted of the classification scores of all the decoy PSMs. For each of the modifications, if there were significant differences (p-value<0.01) between the classification scores of target PSMs and decoy PSMs, and at the same time, between the target PSMs and the universal decoy PSMs, this modification was deemed as validated. Modifications which did not satisfy any of the two conditions were filtered out.

The final step of the multi-stage strategy was performing filtering at all of the three levels: protein, peptide and PSM. A protein inference algorithm was adopted for this step. Then according to the determined protein and peptide scores, the FDRs were controlled at less than 1% for all of the three levels. Usually the protein level FDR was about 1% and the other levels much less than 1%.

According to the filtered PSMs, MS1 peak intensity based label free quantitation was adopted. Then the qualitative and quantitative information of PSMs, peptides, proteins and modifications were exported. For conventional proteomics studies the post analysis workflow ended here. For metaproteomic studies, taxonomy analysis and functional annotation were then performed consecutively.

**Benchmark dataset**

The information of the analyzed benchmark dataset was as follows. For each dataset refer to the corresponding supplementary data for the detailed information about the raw files used, the identified potential modifications, PSMs, peptides and proteins. Some examples of manually
annotated spectra of PSMs with most frequently identified modifications (usually the top 30) were provided for four datasets, including Ecoli_phos (Supplementary Material 1), Mouse_phos (Supplementary Material 2), Hela_phos (Supplementary Material 3), U87_phos (Supplementary Material 4). Examples of manually annotated spectra of glycopeptides were provided for Human_gut dataset (Supplementary Material 5).

<table>
<thead>
<tr>
<th>Dataset ID</th>
<th>Supplementary Data ID</th>
<th>PRIDE id</th>
<th>Species</th>
<th>Tissues/Cell</th>
<th>Enrichment information</th>
<th>Raw file count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo_HEK293</td>
<td>1</td>
<td>PXD001468</td>
<td>Homo sapiens</td>
<td>HEK 293</td>
<td>-</td>
<td>32</td>
</tr>
<tr>
<td>Ecoli_phos</td>
<td>2</td>
<td>PXD008289</td>
<td>Escherichia coli</td>
<td></td>
<td>Phosphorylation</td>
<td>9</td>
</tr>
<tr>
<td>Mouse_phos</td>
<td>3</td>
<td>PXD00792</td>
<td>Mus musculus</td>
<td>Kidney</td>
<td>Phosphorylation</td>
<td>6</td>
</tr>
<tr>
<td>Hela_phos</td>
<td>4</td>
<td>PXD004940</td>
<td>Homo sapiens</td>
<td>HeLa</td>
<td>Phosphorylation</td>
<td>6</td>
</tr>
<tr>
<td>U87_phos</td>
<td>5</td>
<td>PXD009227</td>
<td>Homo sapiens</td>
<td>U-87 glioblastoma cells</td>
<td>Phosphorylation</td>
<td>24</td>
</tr>
<tr>
<td>Mock_micro</td>
<td>6</td>
<td>PXD006118</td>
<td>Mock micro biome</td>
<td>-</td>
<td>-</td>
<td>24</td>
</tr>
<tr>
<td>Mouse_gut</td>
<td>7</td>
<td>PXD003527</td>
<td>Mouse gut microbial community</td>
<td>-</td>
<td>-</td>
<td>32</td>
</tr>
<tr>
<td>Human_gut</td>
<td>8</td>
<td>PXD008870</td>
<td>Human gut microbial community</td>
<td>-</td>
<td>-</td>
<td>55</td>
</tr>
</tbody>
</table>

Functional enrichment analysis

The functional enrichment analysis was performed with STRING (https://string-db.org, version 11.0). For the analysis of the functions of phosphate-ribosylation modified proteins in Ecoli_phos, proteins with above 50 PSM counts were used.
Acknowledgement

This work was supported by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-114), CIHR grant (ECD-144627), the Natural Sciences and Engineering Research Council of Canada (NSERC, grant no. 210034), the Ontario Ministry of Economic Development and Innovation (REG1-4450) and The University of Ottawa.
References


1. Open search result
2. Filtering PSMs based on ΔMass
3. Semi-supervised machine learning based classification
4. Filtering ΔMass peaks by t-test
5. Controlling FDR in PSM, peptide and protein levels
6. MS1 intensity-based label free quantification
7. Taxonomy analysis and functional annotation
* modification on protein n-term which lose the first amino acid Met, so the actual ΔMass is 20.9481 + 131.0405 = 151.9886.

# modification on Cys (fixed modification site by Carbamidomethyl in database search), so the actual ΔMass is 22.945 + 57.0215 = 79.9665.