Microbiome-wide characterization of post-translational modifications

Kai Cheng[1], Zhibin Ning[1], Xu Zhang[1], Janice Mayne[1] and Daniel Figeys[1,2*]

*Correspondence: dfigeys@uottawa.ca

[1]Department 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

[2]Molecular 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.

28 The relationship between the microbiome and human health has been a very active subject of
29 research in recent years[1, 2]. The study of host-microbiome relationships has been primarily focused
30 on the composition of the microbial community and the relative abundance of microbes using DNA
31 sequencing-based technologies, as well as functions of the microbiome using metagenomics and
32 metatranscriptomics[3]. Metaproteomics provides direct assessments of the microbiome functions by
33 qualitative and quantitative investigating the expressed proteins products[4], [5-8]. Post-translational
34 modifications (PTM), which are important for the dynamic regulations of the biological functions of
35 the microbiome, can also be identified by mass spectrometry based metaproteomics[9]. Yet,
36 characterization of the PTM profile from microbial communities bring additional computational
37 challenges in data interpretation. Conventional protein identification strategies allow a few
38 pre-selected PTMs to be identified (variable modifications) but are not suitable to identify multiple
39 modifications and unknown modifications. In contrast, open search strategies allow unbiased
40 identification of potential PTMs and amino acid variants[10]. Unfortunately, the application of open
41 search strategies in metaproteomics has been limited. Open search strategies increase the search
42 space when considering multiple possible mass differences for modified peptides. This is
43 compounded for metaproteomics as protein sequence databases are already very large. Generally,
44 an increase of several hundred-fold in processing time occurs during open searches. Fortunately,
45 several recently developed software tools have significantly improved open search speeds[11, 12],
46 making open search feasible on a metaproteomic scale. Nevertheless, considering the complexity of
47 microbial communities, ensuring the accuracy of peptide and PTM identifications is difficult. Several
48 hundred potential PTMs can be found from one metaproteomic dataset. Calculation of the false
49 discovery rate (FDR) at the peptide spectra match (PSM) level provides only an overall estimation of
50 the confidence of identified peptides.
51 To address this challenge, we developed a multistage filtering strategy for the evaluation of open
52 search results. A series of strict filtering criteria were applied for PTM, PSM, peptide and protein
53 levels, which significantly improved the sensitivity and accuracy of the open search results. Since
54 quantitative analysis is an indispensable step in proteomic data processing workflows, we
55 incorporated a quantitative tool, FlashLFQ[13] into our pipeline. By combining this tool with our
56 previously developed taxonomy analysis and functional annotation modules[14, 15], we built a complete
57 workflow, which enables researchers to study the profile of PTMs of microbiota samples, as well as

58 relationships with taxa, functions and pathways. Moreover, traditional proteomic studies using single
59 species are certainly supported by this strategy. This software is free for academic use and has a
60 user-friendly Graphic User Interface (GUI). Both online and local versions of the tool are available at
61 https://imetalab.ca/.

62 **RESULTS AND DISCUSSION**

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

64 The validation of peptide identification is a tricky problem in proteomics. The most prevalent strategy
65 for validation of methods in proteomic software is the use of a target-decoy database search, which
66 provides an estimate of the probability of random matches. Generally a FDR cutoff at 1% is accepted,
67 however, it has been proposed that a lower FDR threshold (of about 0.1%) is required for open
68 search results, which would ensure a comparable confidence level with the conventional closed
69 search strategy[16]. Moreover, only controlling the FDR at the PSM level is not sufficient for the
70 investigation of PTMs.
71 Herein we developed a multistage filtering strategy for the confident identification of peptides and
72 PTMs using open search (**Figure 1**). Briefly, first the MS/MS spectra were searched against a
73 target-decoy database by open search. Then PSMs were binned according to their mass deviations
74 (ΔMass) (**Figure 2a**). The ΔMass peaks were detected by Gaussian fitting analysis through the whole
75 ΔMass range. The PSMs that did not belong to any ΔMass peaks were filtered out. Then for each
76 Gaussian peak the R^2 was calculated and only peaks with R^2 above 0.9 were retained. The centroid of
77 each peak was also determined during the curve fitting step, which was the experimental mass shift
78 of the potential modification. Noting that the causes for the mass shift are various, including PTM,
79 amino acid substitution, chemical derivative etc., they are all termed as "potential modifications".
80 Usually in closed searches the precursor mass tolerance is set as 10 ppm. Therefore, in this step, the
81 PSMs with ΔMass from the centroid of the corresponding Gaussian peak exceeding 10 ppm were
82 discarded.
83 We first used a previously reported HEK293 data set[10] for demonstration. After the PSM list was
84 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 ΔMass=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 ΔMass 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 ΔMass peaks were kept, among which 160 peaks were identified as monoisotopic peptide peaks (**Online Methods**). As a result, aside from the peak with ΔMass of 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 ΔMass range (about 10% were concentrated in ΔMass=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 ΔMass peak. This result proved that retaining only those PSMs belonging to high quality Gaussian peaks in the ΔMass 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 ΔMass=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 ΔMass Gaussian peak to calculate

115 the FDR of this group, the FDRs were below 1% for only 12 potential modifications (**Supplementary**
116 **Fig. 2**). The FDR of the unmodified peptides which accounted for 80.2% (460,546/574,437) of the
117 total amount of PSMs was only 0.28% (1,290/460,546). This meant that the FDR was higher for the
118 modified portion. Only controlling the FDR at the level of PSM cannot assure the confidence of PSMs
119 with certain modifications. Therefore, we added filtering at the level of modification. Firstly a
120 classification score was assigned to each PSM based on a semi-supervised machine learning
121 strategy[17] (**Online Methods**, **Supplementary Table 1**). Then for each potential modification, an
122 unpaired two-sample t-test was performed between target PSMs and decoy PSMs using the assigned
123 classification score. If the two sets of score values showed significant difference (p-value < 0.01), this
124 modification was retained. Following this step, 17 ΔMass peaks were discarded. The remaining
125 modifications were considered "true" with PSMs having significant differences between target and
126 decoy matches.
127 The last filtering process was adopted at the protein and peptide levels. The protein inference
128 algorithm was utilized at this step to calculate the scores of the proteins and peptides[18]. Then the
129 protein and peptide lists were filtered to ensure that, at both levels, the FDRs were less than 1%. For
130 this dataset the final FDRs were only 0.08% (441/562,302) at the PSM level and 0.15% (179/120,037)
131 at the peptide level, respectively (**Supplementary Data 1**). In total 142 possible modifications were
132 detected, with 50 modifications matched with items in Unimod (http://www.unimod.org/) and 40
133 unknown modifications that can be constructed by combining two modifications from the 50 known
134 modifications (**Supplementary Table 2**). We analyzed the compositions of 40 unknown modifications
135 and found that 72% (29/40) of them contained at least one modification listed below: -156.1019,
136 156.1014, 128.0951 and -128.0958 Da, which corresponded to addition/loss of Arg or Lys. Loss of the
137 terminal amino acid by in-source fragmentation was common in ESI-MS. Addition of Arg or Lys could
138 be caused by additional missed cleavage sites (consecutive Arg or Lys in the peptide C-term). In
139 addition, 25% (10/40) of the combinations contained -71.0375 Da (Gln->Gly), -71.0743 Da (Lys->Gly)
140 or -57.0221 Da (Asn->Gly/Gln->Ala), which were related to amino acid substitutions. Therefore, from
141 the combination of information we can determine that potential modifications without Unimod
142 matched items, can be explained by known modifications to a large extent.
143 In summary, we applied this method to a large scale proteomic dataset from the HEK293 cell line.
144 The results illustrated that high confident identifications at all of the four levels (PSMs, peptides,

145 proteins and potential modifications) were obtained. Compared with conventional database search
146 methods, the sensitivity was significantly improved by our strategy (**Supplementary Fig. 3a**). Using
147 this strategy 562,302 PSMs were identified with PSM level FDR<0.08%, and peptide and protein level
148 FDR<1%. In a closed search in which even the peptide and protein level FDR were not considered,
149 434,526 PSMs were obtained with PSM level FDR<1%. If we continued to restrict the FDR at the
150 peptide and protein levels to 1%, 385,990 PSMs remain (**Supplementary Fig. 3b**). The significant
151 improvement of sensitivity using our strategy was not only because of the identification of modified
152 peptides, which cannot be identified in the conventional method; in fact, the 459,367 unmodified
153 PSMs obtained in open search were more than the number obtained in closed search. With proper
154 filtering, the increased search space did not sacrifice the identification sensitivity; on the contrary, it
155 improved it. The first reason was 36.2% (16,969/46,868) decoy PSMs were matched to more credible
156 targets. This portion of loss was recovered in open search, at the same time the additional expense
157 for the elimination of these random matches in FDR control was avoided. The second reason was
158 that random matches were basically evenly distributed over a wider range, which made the ΔMass a
159 significant feature to distinguish random matches effectively. Based on these reasons we developed
160 the post-analysis workflow for the analysis of open search datasets, which showed greatly improved
161 accuracy and sensitivity of identifications.

162 **Identification of a group of phosphorylation related modifications from enriched samples**

163 To further benchmark the post-validation strategy, we re-analyzed *E. coli* phospho-enriched samples
164 (Dataset ID: Ecoli_phos)[19] to measure the proportion of PSMs that could be identified as
165 phosphopeptides. Authenticating our strategy, phosphorylation was identified as the most abundant
166 modification from the phospho-enriched *E. coli* sample. Even more substantiating, the 2nd-4th
167 potential modifications were phosphate-ribosylation (ΔMass=212.0091), PhosphoHex
168 (ΔMass=242.0195) and 3-phosphoglyceryl (ΔMass=167.983), respectively, which were all
169 phosphorylation related modifications. Actually, by manual interpretation of the spectra from the top
170 30 abundant mass differences, we found that neutral loss of phosphorylation was observed from 14
171 types of potential modifications (**Supplementary Data 2, Supplementary Material 1**). Aside from
172 phosphorylation, pyrophosphorylation (ΔMass=159.9328), and the abundant phospho-related

173 modifications mentioned above, other known phospho-related modifications also include
174 phosphogluconoylation (ΔMass=258.0149) and pyridoxal phosphate (ΔMass=229.0134). No Unimod
175 records were found for possible modifications with experimental ΔMass 151.9878. We inferred the
176 composition as H(5)C(3)O(5)P because according to the ΔMass it is most likely deducting one oxygen
177 from 3-phosphoglyceryl. We also inferred the composition of ΔMass 181.9985 as H(7)C(4)O(6)P(1),
178 because the mass difference between this and 3-phosphoglyceryl is 14.0154 Da which could be a
179 methylene group. The possible structures of the identified phospho-related modifications were
180 illustrated (**Figure 3a**).
181 Phosphate-ribosylation is a tag of ADP-Ribosylation, which is a versatile PTM involved in various
182 cellular processes[20, 21]. It has been reported that ADP-Ribosylation sites could be detected in
183 phosphoproteomics data by setting it as the corresponding variable modification during database
184 searches[22]. In this dataset phosphate-ribosylation were identified as the 2nd most frequent PSM
185 count. In theory, the ADP-ribose group can be attached on multiple amino acids, such as Asp, Glu, Lys,
186 Arg and Cys. The specificity sites for phosphate-ribosylation recorded in Unimod are Asp, Glu and Arg.
187 However, in this dataset, the modification sites were observed as Lys and protein N-term. In total,
188 86.3% (864/1001) of phosphate-ribosylation modified peptides contain a missed cleavage site Lys.
189 Additionally, 7.4% (74/1001) of peptides start with the first amino acid of the proteins, and 8.9%
190 (89/1001) of peptides start with the second amino acid of the proteins, for a total of 16.3%
191 (163/1029) peptides containing protein N-term (**Supplementary Material 1**). Through the functional
192 enrichment analysis of the proteins with phosphate-ribosylation sites, 61 biological processes with
193 FDR below 0.05 were obtained (**Figure 3b, Supplementary Table 3, 4**). The colored three biological
194 processes were all branches of nucleoside phosphate metabolic process, which suggested the
195 important role of ADP-ribosylation in this process. The 3rd and 4th potential modifications were
196 matched with PhosphoHex and 3-phosphoglyceryl, respectively. As above, most phosphoHex and
197 3-phosphoglyceryl modified peptides contained a missed cleavage site (Lys) or protein N-term. The
198 manually annotated spectra showed that the modifications occurred on the Lys at the missed
199 cleavage site or protein N-term (**Supplementary material 1**). These results clearly showed the
200 advantage of open search for the identification of coexisting PTMs from enriched samples. For
201 example, Lys and protein N-term were not specified sites of phosphate-ribosylation and phosphoHex
202 in Unimod, so these sites would be missed when setting these as variable modifications in closed

203 searches. However, in open searches, novel modifications or known modifications on novel sites can
204 be readily identified. The proteome-wide investigation of phosphate-ribosylation sites and their
205 abundance is important for the further understanding of ADP-ribosylation functions.
206 Encouraged by this success we investigated phosphorylation enriched mammalian samples from
207 mouse kidney (Dataset ID: Mouse_phos)[23]. Surprisingly, only two of the top 30 ΔMasses were
208 matched with known modifications in Unimod, namely phosphorylation and pyrophosphorylation
209 (**Supplementary Data 3**). It's easy to find ΔMasses 239.8978 and 319.8637 that represented multiple
210 phosphorylation sites occurring on one peptide. By manually interpretation of the spectra we found
211 that all of the unknown ΔMasses were combinations of multiple modifications, and in most cases,
212 were combinations of phosphorylation and others (**Supplementary material 2**). In fact, among the
213 top 30 abundant ΔMasses, 26 contained at least one phosphate group, which strongly confirmed the
214 confidence of this result. Four of them were combinations of phosphorylation and cations (protons
215 replaced by Na^+ or K^+). This type was not observed for *Ecoli*_phos (but common in mouse_phos).
216 While 3-phosphoglyceryl and phosphate-ribosylation were found with few PSMs counts and
217 phosphoHex was not observed. Apart from this, peptides from mouse_phos tended to contain
218 multiple modifications, which was a significant difference between these two datasets.
219 We also studied two human cell lines, HeLa[24] (Dataset ID: Hela_phos, **Supplementary Data 4,**
220 **Supplementary material 3**) and U-87 glioblastoma cells (Dataset ID: U87_phos, **Supplementary Data**
221 **5, Supplementary material 4**)[25]. Similar to the mouse_phos dataset, multiple phosphorylation and
222 combinations of phosphorylation and other modifications were prevalent. But complex
223 phospho-related modifications as these existed in *Ecoli*_phos data were very rarely observed. Our
224 results clearly showed the different features of phosphopeptides identified from bacteria and
225 mammal cells. In *Ecoli*_phos the majority of peptides have only one modification, including some
226 complex modification forms and unknown modifications. From the mammalian datasets, multiple
227 phosphorylation sites on one peptide were common rather than phospho-related complex
228 modifications. At the same time, peptides modified by cations were often observed from mammalian
229 datasets, but rare in *Ecoli*_phos datasets. Using our strategy co-enriched modifications can be
230 identified in one step which is of great significance for the analysis of interactions of different
231 modifications. Collectively, these results showed significant benefits of the application of open
232 searches for enriched sample analyses.

233 **Identifying the modification profile from microbiome samples**

234 We also applied our workflow to analyze a microbiome dataset which was constructed by assembling
235 32 species/strains of Archaea, Bacteria, Eukaryotes and Bacteriophages[26] (**Supplementary Note,**
236 **Supplementary Data 6**). Given the advantage of known microbial composition in this dataset, we
237 also used this dataset to benchmark the quantitative accuracy of our method, to investigate if the
238 quantitation will be affected by the additionally identified peptides from open search. In total, we
239 identified 1,048,759 PSMs, which was 16% more than from conventional closed searches
240 (**Supplementary Fig. 4**). Through quantitative analysis of the mock microbiota community,
241 satisfactory quantitative accuracy was obtained at the species-level, which also confirmed the
242 confidence of the identifications of modified peptides (**Supplementary Fig. 5**, **Supplementary Note**).
243 To demonstrate the capability of open searches for microbiome sample analyses we applied the
244 developed strategy to analyze a mouse gut microbiome dataset (Dataset ID: Mouse_gut)[27]. In total
245 918,132 PSMs were identified with an estimated FDR of 0.17%. The PSM count increased
246 significantly by 28.3%, 202,682 additional PSMs were identified when compared with our previous
247 work [14]. At the peptide level the FDR is 0.33% and 127,218 unique sequences (37,544 peptides
248 contained at least one modified form) were obtained (**Supplementary Data 7**). In closed search
249 115,832 peptide identifications were obtained and so the open search strategy resulted in a 9.8%
250 increase in peptide identifications (**Figure 4a**). For the commonly identified portion, most of the
251 peptides were only identified as unmodified in the open search, but there were about 20% peptides
252 that contained at least one modified form. For the additional identified portion by open search,
253 about 40% of peptides were found only in modified forms and over 50% were unmodified. This result
254 showed that, not only was the missing information of modified peptides retrieved from the open
255 search, but also, that the improved sensitivity of the workflow enabled the additional detection of
256 unmodified peptides. The increase in the number of confidently recognized species (with at least
257 three unique peptides) was also realized. In this work 172 species were identified containing equal or
258 more than three unique peptides. The number was 143 from the previous work, with a major
259 common portion of 134 species shared by the two results (**Figure 4b**). The 29 species identified only
260 in this workflow were illustrated and their distribution at the phylum level was shown in **Figure 4b**.
261 Most of the newly found species were from phyla *Firmicutes* and *Bacteroidetes*. Actually, 27 species

262 were also identified using the conventional method but without enough unique peptide counts,
263 which also showed the confidence of the species specifically determined using the new strategy.
264 From this result we can conclude that through our strategy, comparable accuracy and optimized
265 sensitivity can be obtained by open searches when compared with traditional closed searches for the
266 analysis of real microbiota samples.

267 In this experiment the mouse gut microbiome were collected from both high-fat diet (HFD) and
268 low-fat diet (LFD) fed mice for the interpretation of diet-microbiota-host interactions[27]. In this work,
269 based on the quantitative information of the potential modifications, we investigated the response
270 of the determined modifications to diet in mouse microbiomes. The abundances of various
271 modifications were calculated in each identified taxon node. Both the absolute and relative values
272 were considered, that is, for each taxon the summation of the intensity of peptides with specific
273 modifications were calculated as the absolute abundance of the modification in this taxon, and the
274 ratio between the absolute value and the total abundance of this taxon were calculated as the
275 relative abundance. Firstly, we tested if there were modifications that showed significant differences
276 between HFD and LFD. We found that the modification ΔMass=71.0368 (Acrylamide adduct/Gly->Gln
277 substitution) showed significant increases in the HFD samples using both absolute and relative values,
278 and the modification ΔMass=52.1229 (replacement of 3 protons by iron) decreased significantly in
279 relative value (**Figure 5a, Supplementary material 5**). This result confirmed the ability of our strategy
280 for the identification of quantitative changes of PTMs from real gut microbiome samples.

281 Since there were two possibilities for the modification ΔMass=71.0368, we investigated the
282 information of localization sites to determine which was validate. Acrylamide adducts may occur on
283 Cys, Lys and any amino acid on peptide N-term whereas Gly->Gln substitution only exists on Gly.
284 Considering all the possible sites for localization analysis, we found Lys the most frequent
285 modification site for ΔMass=71.0368 (**Supplementary Fig.6**), which suggested that the most
286 probable modification was an acrylamide adduct. The relative abundance of modified peptides in
287 each taxonomy in both HFD and LFD samples were illustrated respectively (**Figure 5b**). The
288 compositions and relative abundances were almost constant in these two conditions. The lineage
289 *Bacteria->Firmicutes->Bacilli->Lactobacillales->Streptococcaceae->Lactococcus->Lactococcus lactis*
290 was the main source for acrylamide adduct modifications and contributed to the high abundance in
291 HFD samples directly. At the protein level the difference was derived from the single site K264 from

292 protein glyceraldehyde-3-phosphate dehydrogenase (GAPDH) in *Lactococcus lactis*. GAPDH is
293 involved in the glucose metabolic process and related to the function NAD binding, NADP binding
294 and GAPDH (NAD+) (phosphorylating) activity. Acrylamide adducts have been deemed as potentially
295 toxic compounds in food[28]. The degradation of acrylamide by bacteria was widely studied[29, 30] and
296 *Bacillus cereus* has been reported to efficiently degrade acrylamide[31]. Molecular mechanisms of the
297 addition of acrylamide adducts on GAPDH was studied by incubating purified human GAPDH with
298 acrylamide and the kinetics of Cys acrylamide adduct formation was characterized, but the
299 modification on Lys was not described[32]. This work presented the novel discovery of Lys acrylamide
300 adducts on GAPDH from *Lactococcus lactis*, and the potential relationship with HFDs.
301 For the modification replacement of three protons by iron (cation:Fe[III] for short), the most
302 abundant species were *Bacteroides vulgatus* and *Parabacteroides goldsteinii* from the same order
303 *Bacteroidales* (**Figure 5c**). Significant change was only observed in species *Parabacteroides goldsteinii*.
304 The involved proteins included TonB-dependent receptor, RagB/SusD family nutrient uptake outer
305 membrane protein, SusC/RagA family TonB-linked outer membrane protein and hypothetical
306 proteins. The first three were all membrane proteins and TonB-dependent receptor was a known iron
307 transporter[33]. The identification of cation:Fe[III] modification on these proteins could help
308 researchers further their understanding of the transport mechanism of bacteria.

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

310 We then applied this workflow for the analysis of a human gut microbiome dataset from a study
311 about type 1 diabetes (Dataset ID: Human_gut)[34]. Two subject groups of data from healthy control
312 (CO, n=22) and new-onset patients (NO, n=33) were analyzed. From this dataset we identified
313 829,458 PSMs at FDR about 0.16% (**Supplementary Data 8**). Only 676,282 PSMs were obtained at
314 the FDR about 4.1% by the conventional method (**Supplementary Fig. 7**). The identification rate was
315 increased from 35.8% to 44.0%, and while a much lower FDR was obtained. For each sample, 10,305
316 ± 2,205 (mean ± standard deviation) unique peptide sequences and 4,763 ± 1,027 proteins were
317 quantified. In total 43,475 non-human microbiome proteins and 629 human proteins were identified.
318 We quantified 611 species, among which 240 species were identified with more than three unique
319 peptides. Annotations with KEGG database allowed us to identify 459 ± 69 microbial and 100 ± 27

320 human KEGG orthologies (KOs) per sample.

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

328 Currently the glycosylation in the microbial community is an under-studied area. For a better
329 understanding of the gut microbiome-host interactions the characterization of glycosylation is
330 absolutely necessary. In this study, five types of glycosylation including Hex (hexose, ΔMass =
331 162.053), HexNAc (N-Acetylhexosamine, ΔMass = 203.079), HexNAc1dHex1 (ΔMass = 349.137),
332 HexNAc2 (ΔMass = 406.159) and bacillosamine (2,4-diacetamido-2,4,6-trideoxyglucopyranose,
333 ΔMass = 228.1110) were identified from this dataset (**Figure 6a**). An important feature of the
334 glycopeptide spectra were that the diagnostic ions could be observed in low mass-to-charge ratios
335 (m/z), which were called glycan oxonium ions[35]. In our results, we detected at least one glycan
336 oxonium ion whose intensity exceeded 50% of the peaks in the corresponding MS2 spectra, in 92.0%
337 (5,424/5,895), 94.6% (5,750/6,078), 93.4% (157/168) and 69.5% (91/131) of the HexNAc,
338 HexNAc1dHex1, HexNAc2, and bacillosamine modified PSMs, respectively (**Supplementary Table 5**).
339 Manual verifications were also adopted for part of the glycopeptide spectra, and generally
340 successive y-ions were observed in these spectra (**Supplementary material 6**). These results ensured
341 the reliability of the identification of glycopeptides. Interestingly, we found the distributions of these
342 glycosylations between human and microbial proteins differed. Two types of glycosylation with
343 apparently higher abundances, HexNAc1dHex1 and HexNAc were only observed on human proteins.
344 For Hex and HexNAc2, there were 38.1% (45/118) and 28.6% (48/168) of the PSMs that were
345 identified from microbial proteins, respectively. Bacillosamine was a rare monosaccharide which was
346 found in bacterial glycoproteins[36]. And in our results 93.1% (122/131) of the bacillosamine modified
347 PSMs were from bacteria proteins.

348 Glycosylation plays a critical role in the regulation of protein functions. The alterations of
349 glycosylation could affect the immune system of mammals[37]. The relationship between abnormal

350 glycosylation and the pathogenesis of diabetes has also been reported[38]. In this dataset, significant
351 decreases in the abundance of HexNAc modified peptides were detected in NO patients compared
352 to CO subjects (**Figure 6b**). Functional enrichment analysis of HexNAc modified proteins was
353 performed (**Figure 6c, Supplementary Table 6**). Neutrophil related immune response was the most
354 significantly enriched biological processes (**Supplementary Fig. 10**). Among the enriched Gene
355 Ontology (GO) terms, nine of them were significantly changed (**Supplementary Fig. 11**). Five of them
356 (transport, response to stress, inflammatory response, response to stimulus and localization) were
357 also downregulated in NO samples which were consistent with the general trend. But four catabolic
358 processes (catabolic process, cellular catabolic process, organonitrogen compound catabolic process
359 and organic substance catabolic process) related functions were upregulated.
360 Besides high abundant human protein glycosylation, we also identified three types of glycosylation
361 from microbial proteins. For hexose, 21 glycopeptides were identified, only one peptide was from
362 *Bacteroidetes* and 16 peptides were from *Firmicutes* (**Supplementary Fig. 12, Supplementary table**
363 **7**). In this dataset, 86.7% (39/45) of the hexose modified PSMs were from *Clostridiales*. From the
364 functional annotation, we found that the most frequently observed function related to hexose was
365 ABC-type glycerol-3-phosphate transport system, periplasmic component (**Figure 6d**). For HexNAc2,
366 one peptide EAATTAAAAAATTDVADKK was identified from eight CO and seven NO samples,
367 respectively, which was the most widely identified glycopeptide from bacteria in Human_gut
368 (**Supplementary Fig. 13, Supplementary table 7**). The lowest common ancestor (LCA) of this peptide
369 was *Clostridia* (order level). The corresponding protein was ABC-type sugar transport system,
370 periplasmic component, contains N-terminal xre family HTH domain, which was also an ATP-binding
371 cassette (ABC) transporter. It could be seen that the ABC transporter families were the most
372 frequently identified glycoproteins from bacteria in Human_gut. We also identified bacteria specific
373 monosaccharide Bacillosamine from 48 glycopeptides. Bacillosamine was described as an
374 "Asn-linked glycan from Gram-negative Bacterium" in Unimod. However, we noticed that only 45.8%
375 (22/48) glycopeptides contained Asn, which suggested the existence of other types such as O-linked
376 glycopeptides in these samples. Bacillosamine was involved in various functions in this dataset
377 (**Supplementary Table 7**). We identified the LCA for 21 glycopeptides (**Supplementary Fig. 14**). Three
378 peptides determined in superkingdom level as bacteria were from the protein Elongation factor Tu
379 (EF-Tu), which was a highly conserved proteins in prokaryotes. EF-Tu was a type of GTPase, which

380 was also an important antibiotic target[39]. For the remainder of the glycopeptides, seven were from
381 Bacteroidetes and eight were from Firmicutes. The main COG categories were carbohydrate
382 transport and metabolism (G) and energy production and conversion (C). The characterization of
383 bacteria glycosylation will greatly improve our understanding of gut microbiome-host interactions.

384 **Conclusion**

385 Modifications of peptides bring mass variations which can increase the number of unidentified
386 spectra in traditional database searches. The open search strategy is a method capable of identifying
387 modifications in an unrestricted manner, however, the validation of the results remains challenging,
388 especially for the studies of microbial communities. To address this, we developed a multistage
389 filtering strategy, which could ensure the confidence of the results at various levels: PSMs, peptides,
390 proteins and modifications. In this method, rigorous filters are utilized to remove potential
391 modifications with less confidence. The analysis of the benchmark datasets (Homo_HEK293,
392 Mock_micro, Mouse_gut and Human_gut) suggests that high accuracy is achieved without the
393 sacrifice of sensitivity. It is worth noting that the identification rates are greatly improved in the
394 analysis of complex mouse/human gut microbiota samples using open search strategy. Combined
395 with the previously developed metaproteomics data analysis tools by our group, a novel
396 bioinformatics workflow is provided for the comprehensive analysis of PTM profiles from microbial
397 communities. Rich and quantitative information can be obtained easily and with high accuracy, which
398 is the major reason we recommend this strategy to be a standard workflow for
399 proteomics/metaproteomics studies.
400

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

434 mouse gut microbiota samples.

435 Figure 5.

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

440 Figure 6.

441 **(a)** Identified potential modifications and their distribution between human and microbial proteins.
442 Purple dots: modifications matched in Unimod; Grey dots: unknown modifications. The annotated
443 modifications fulfilled one of the following two conditions: 1) with above 1,000 PSM counts; 2) with
444 above 100 PSM counts and more than 90% were from microbial proteins. The possible structures of
445 five glycosylation were illustrated. **(b)** HexNAc modified glycopeptides showed significant difference
446 between CO and NO samples. **(c)** The enriched biological process of HexNAc modified human
447 glycoproteins. **(d-e)** The functions of the microbial proteins modified by **(d)** Hex and **(e)** HexNAc. CO:
448 control subjects; NO; new-onset patients.
449

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 ΔMass. Through the ΔMass range, the Gaussian peaks with R^2 above 0.9 were detected (users can change this parameter in our software). PSMs don't belong to any ΔMass 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[40]. When we found two Gaussian peaks, $Peak_i$ ($\Delta Mass=M_i$) and $Peak_j$ ($\Delta Mass=M_j$), which fulfilled the conditions that $|M_j-M_i-1.00286864|\leq 0.01$, 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[18]. 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[13]. 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[14, 15].

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**).

Dataset ID	Supplementary Data ID	PRIDE id	Species	Tissues/Cell	Enrichment information	Raw file count
Homo_HEK293[10]	1	PXD001468	Homo sapiens	HEK293	-	32
Ecoli_phos[19]	2	PXD008289	Escherichia coli	-	Phosphorylation	9
Mouse_phos[23]	3	PXD001792	Mus musculus	Kidney	Phosphorylation	6
Hela_phos[24]	4	PXD004940	Homo sapiens	HeLa	Phosphorylation	6
U87_phos[25]	5	PXD009227	Homo sapiens	U-87 glioblastoma cells	Phosphorylation	24
Mock_micro[26]	6	PXD006118	Mock microbiome	-	-	24
Mouse_gut[27]	7	PXD003527	Mouse gut microbial community	-	-	32
Human_gut[34]	8	PXD008870	Human gut microbial community	-	-	55

Functional enrichment analysis

The functional enrichment analysis was performed with STRING (https://string-db.org, version 11.0)[41]. 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.

520 # References

521 1. Knight, R. et al. The Microbiome and Human Biology. *Annu Rev Genom Hum G* 18, 65-86 (2017).

522 2. Gilbert, J.A. et al. Current understanding of the human microbiome. *Nat Med* 24, 392-400 (2018).

523 3. Knight, R. et al. Best practices for analysing microbiomes. *Nat Rev Microbiol* 16, 410-422 (2018).

524 4. White, R.A., Callister, S.J., Moore, R.J., Baker, E.S. & Jansson, J.K. The past, present and future of
525 microbiome analyses. *Nat. Protoc.* 11, 4-8 (2016).

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

528 6. Xiong, W.L., Giannone, R.J., Morowitz, M.J., Banfield, J.F. & Hettich, R.L. Development of an
529 Enhanced Metaproteomic Approach for Deepening the Microbiome Characterization of the Human
530 Infant Gut. *J Proteome Res* 14, 133-141 (2015).

531 7. Maier, T.V. et al. Impact of Dietary Resistant Starch on the Human Gut Microbiome, Metaproteome,
532 and Metabolome. *Mbio* 8 (2017).

533 8. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases.
534 *Nature* 569, 655-662 (2019).

535 9. Zhang, W.P. et al. Post-translational modifications are enriched within protein functional groups
536 important to bacterial adaptation within a deep-sea hydrothermal vent environment. *Microbiome* 4
537 (2016).

538 10. Chick, J.M. et al. A mass-tolerant database search identifies a large proportion of unassigned
539 spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* 33, 743-749 (2015).

540 11. Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D. & Nesvizhskii, A.I. MSFragger:
541 ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat*
542 *Methods* 14, 513-520 (2017).

543 12. Chi, H. et al. Comprehensive identification of peptides in tandem mass spectra using an efficient
544 open search engine. *Nat. Biotechnol.* 36, 1059-+ (2018).

545 13. Millikin, R.J., Solntsev, S.K., Shortreed, M.R. & Smith, L.M. Ultrafast Peptide Label-Free
546 Quantification with FlashLFQ. *J Proteome Res* 17, 386-391 (2018).

547 **14.** Cheng, K. et al. MetaLab: an automated pipeline for metaproteomic data analysis. *Microbiome* 5,
548 157 (2017).

549 **15.** Liao, B. et al. iMetaLab 1.0: a web platform for metaproteomics data analysis. *Bioinformatics* 34,
550 3954-3956 (2018).

551 **16.** Li, D.H. et al. Optimal Settings of Mass Spectrometry Open Search Strategy for Higher Confidence. *J*
552 *Proteome Res* 17, 3719-3729 (2018).

553 **17.** Kall, L., Canterbury, J.D., Weston, J., Noble, W.S. & MacCoss, M.J. Semi-supervised learning for
554 peptide identification from shotgun proteomics datasets. *Nat. Methods* 4, 923-925 (2007).

555 **18.** Uszkoreit, J. et al. PIA: An Intuitive Protein Inference Engine with a Web-Based User Interface. *J*
556 *Proteome Res* 14, 2988-2997 (2015).

557 **19.** Potel, C.M., Lin, M.H., Heck, A.J.R. & Lemeer, S. Defeating major contaminants in Fe(3+)-IMAC
558 phosphopeptide enrichment. *Mol Cell Proteomics* (2018).

559 **20.** Bock, F.J., Todorova, T.T. & Chang, P. The Promise of Proteomics for the Study of ADP-Ribosylation.
560 *Mol Cell* 58, 911-924 (2015).

561 **21.** Luscher, B. et al. ADP-Ribosylation, a Multifaceted Posttranslational Modification Involved in the
562 Control of Cell Physiology in Health and Disease. *Chem. Rev.* 118, 216-260 (2018).

563 **22.** Matic, I., Ahel, I. & Hay, R.T. Reanalysis of phosphoproteomics data uncovers ADP-ribosylation sites.
564 *Nat. Methods* 9, 771-772 (2012).

565 **23.** Humphrey, S.J., Azimifar, S.B. & Mann, M. High-throughput phosphoproteomics reveals in vivo
566 insulin signaling dynamics. *Nat. Biotechnol.* 33, 990-U142 (2015).

567 **24.** Espadas, G., Borras, E., Chiva, C. & Sabido, E. Evaluation of different peptide fragmentation types
568 and mass analyzers in data-dependent methods using an Orbitrap Fusion Lumos Tribrid mass
569 spectrometer. *Proteomics* 17 (2017).

570 **25.** Humphrey, S.J., Karayel, O., James, D.E. & Mann, M. High-throughput and high-sensitivity
571 phosphoproteomics with the EasyPhos platform. *Nat. Protoc.* 13 (2018).

572 **26.** Kleiner, M. et al. Assessing species biomass contributions in microbial communities via
573 metaproteomics. *Nat Commun* 8 (2017).

574 **27.** Zhang, X. et al. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut
575 microbiota. *Microbiome* 4 (2016).

576 **28.** Friedman, M. Acrylamide: inhibition of formation in processed food and mitigation of toxicity in
577 cells, animals, and humans. *Food Funct* 6, 1752-1772 (2015).

578 **29.** Nawaz, M.S., Franklin, W. & Cerniglia, C.E. Degradation of Acrylamide by Immobilized Cells of a
579 Pseudomonas Sp and Xanthomonas-Maltophilia. *Can J Microbiol* 39, 207-212 (1993).

580 **30.** Wyatt, J.M. & Knowles, C.J. Microbial-Degradation of Acrylonitrile Waste Effluents - the
581 Degradation of Effluents and Condensates from the Manufacture of Acrylonitrile. *Int Biodeter*
582 *Biodegr* 35, 227-248 (1995).

583 **31.** Shukor, M.Y. et al. Isolation and characterization of an acrylamide-degrading Bacillus cereus. *J*
584 *Environ Biol* 30, 57-64 (2009).

585 **32.** Martyniuk, C.J. et al. Molecular Mechanism of Glyceraldehyde-3-phosphate Dehydrogenase
586 Inactivation by alpha,beta-Unsaturated Carbonyl Derivatives. *Chem Res Toxicol* 24, 2302-2311
587 (2011).

588 **33.** Noinaj, N., Guillier, M., Barnard, T.J. & Buchanan, S.K. TonB-Dependent Transporters: Regulation,
589 Structure, and Function. *Annual Review of Microbiology, Vol 64, 2010* 64, 43-60 (2010).

590 **34.** Gavin, P.G. et al. Intestinal Metaproteomics Reveals Host-Microbiota Interactions in Subjects at Risk
591 for Type 1 Diabetes. *Diabetes Care* 41, 2178-2186 (2018).

592 **35.** Eshghi, S.T. et al. Classification of Tandem Mass Spectra for Identification of N- and O-linked
593 Glycopeptides. *Sci Rep-Uk* 6 (2016).

594 **36.** Young, N.M. et al. Structure of the N-linked glycan present on multiple glycoproteins in the
595 gram-negative bacterium, Campylobacter jejuni. *J Biol Chem* 277, 42530-42539 (2002).

596 **37.** Marth, J.D. & Grewal, P.K. Mammalian glycosylation in immunity. *Nat Rev Immunol* 8, 874-887
597 (2008).

598 **38.** Bermingham, M.L. et al. N-Glycan Profile and Kidney Disease in Type 1 Diabetes. *Diabetes Care* 41,
599 79-87 (2018).

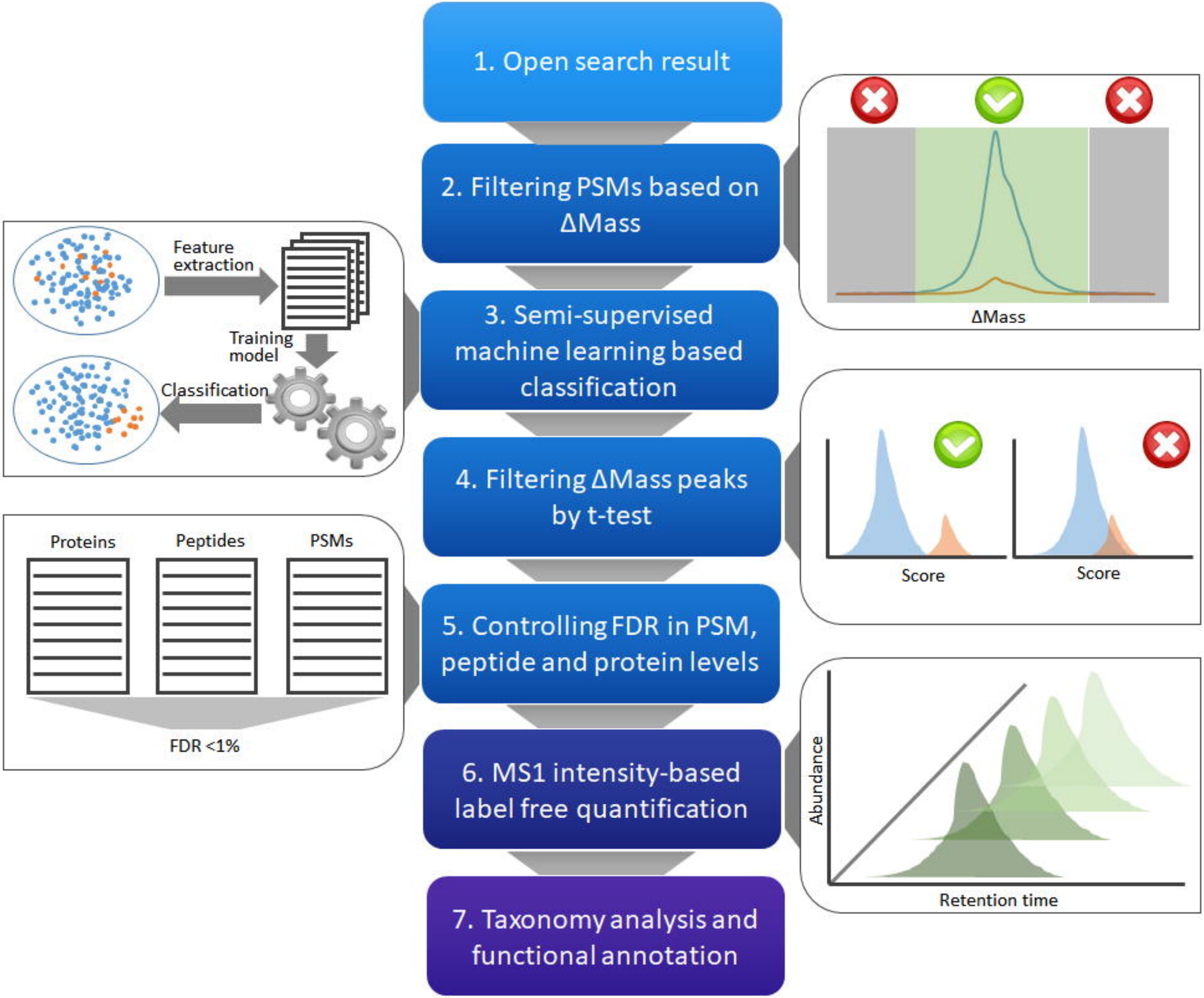
600 **39.** Parmeggiani, A. & Nissen, P. Elongation factor Tu-targeted antibiotics: Four different structures,
601 two mechanisms of action. *Febs Lett* 580, 4576-4581 (2006).

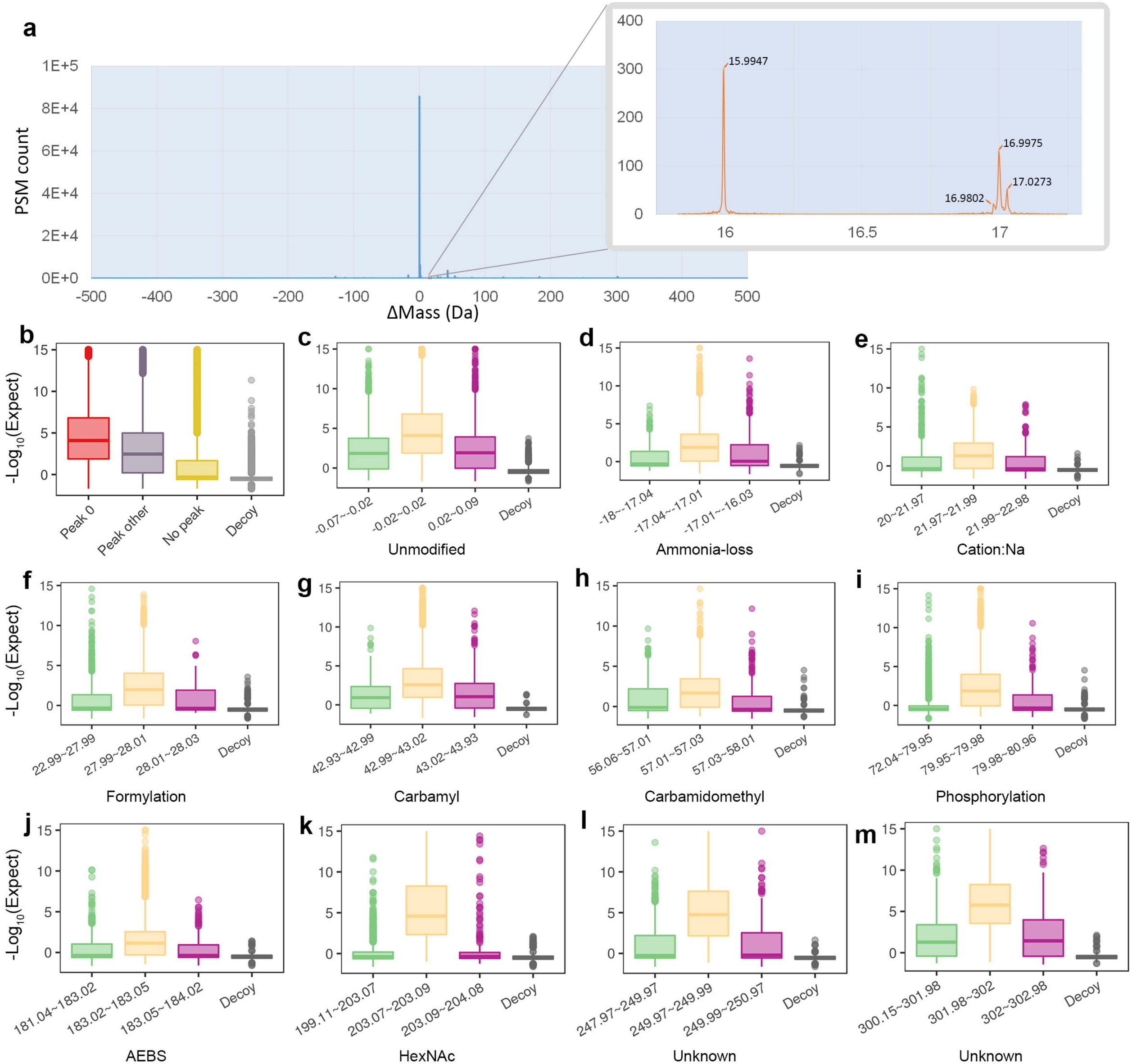
602 **40.** Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range
603 mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 26, 1367-1372 (2008).

604 **41.** Szklarczyk, D. et al. STRING v10: protein-protein interaction networks, integrated over the tree of
605 life. *Nucleic Acids Res.* 43, D447-D452 (2015).

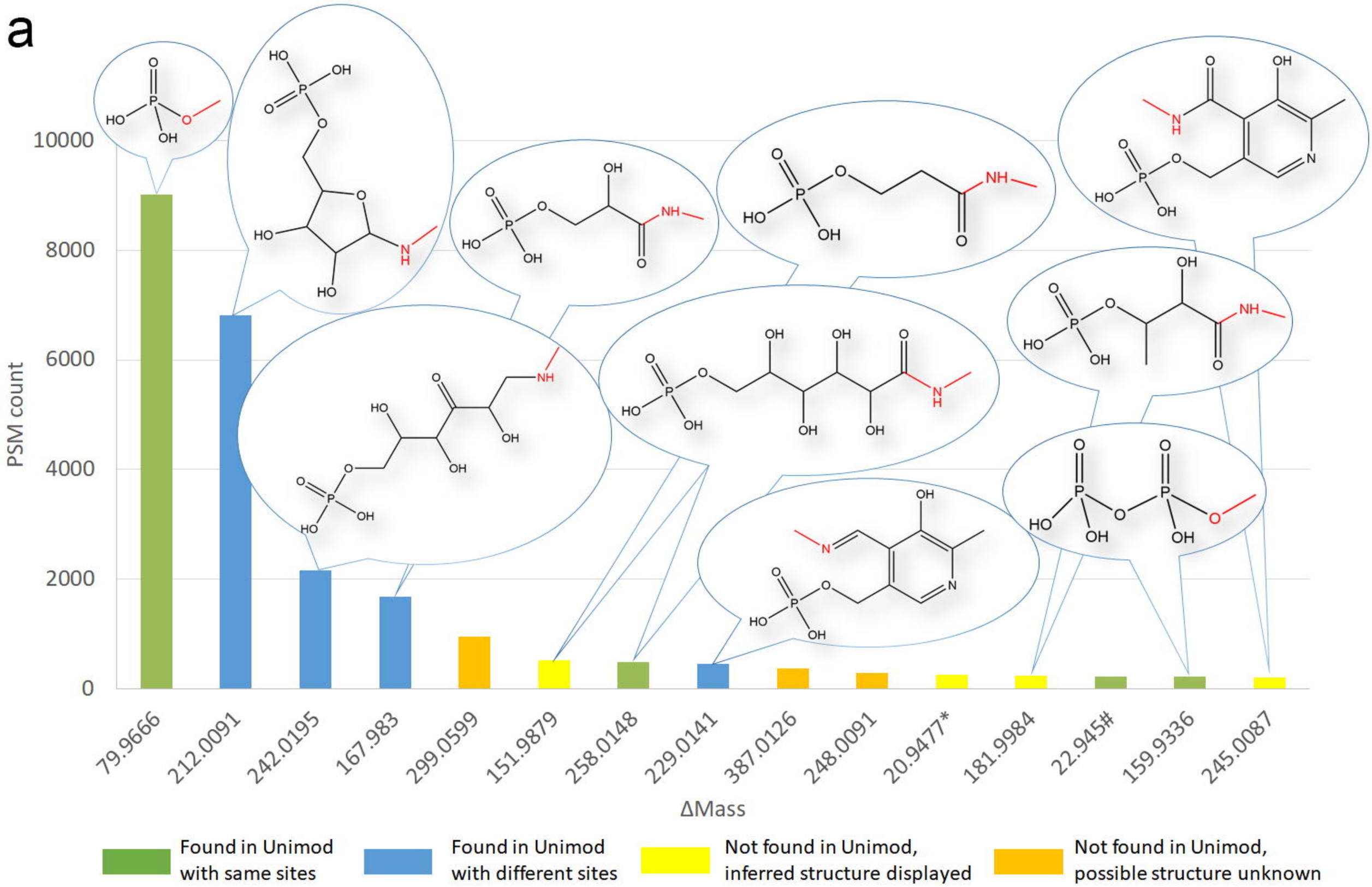
606

607


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
Feature extraction
Training model
Classification
Proteins
Peptides
PSMs
FDR <1%
ΔMass
Score
Score
Abundance
Retention time


a
PSM count
ΔMass (Da)
15.9947
16.9975
17.0273
16.9802
b
Peak 0
Peak other
No peak
Decoy
c
-0.07~-0.02
-0.02~0.02
0.02~0.09
Decoy
Unmodified
d
-18~-17.04
-17.04~-17.01
-17.01~-16.03
Decoy
Ammonia-loss
e
20~21.97
21.97~21.99
21.99~22.98
Decoy
Cation:Na
f
22.99~27.99
27.99~28.01
28.01~28.03
Decoy
Formylation
g
42.93~42.99
42.99~43.02
43.02~43.93
Decoy
Carbamyl
h
56.06~57.01
57.01~57.03
57.03~58.01
Decoy
Carbamidomethyl
i
72.04~79.95
79.95~79.98
79.98~80.96
Decoy
Phosphorylation
j
181.04~183.02
183.02~183.05
183.05~184.02
Decoy
AEBS
k
199.11~203.07
203.07~203.09
203.09~204.08
Decoy
HexNAc
l
247.97~249.97
249.97~249.99
249.99~250.97
Decoy
Unknown
m
300.15~301.98
301.98~302
302~302.98
Decoy
Unknown
-Log10(Expect)

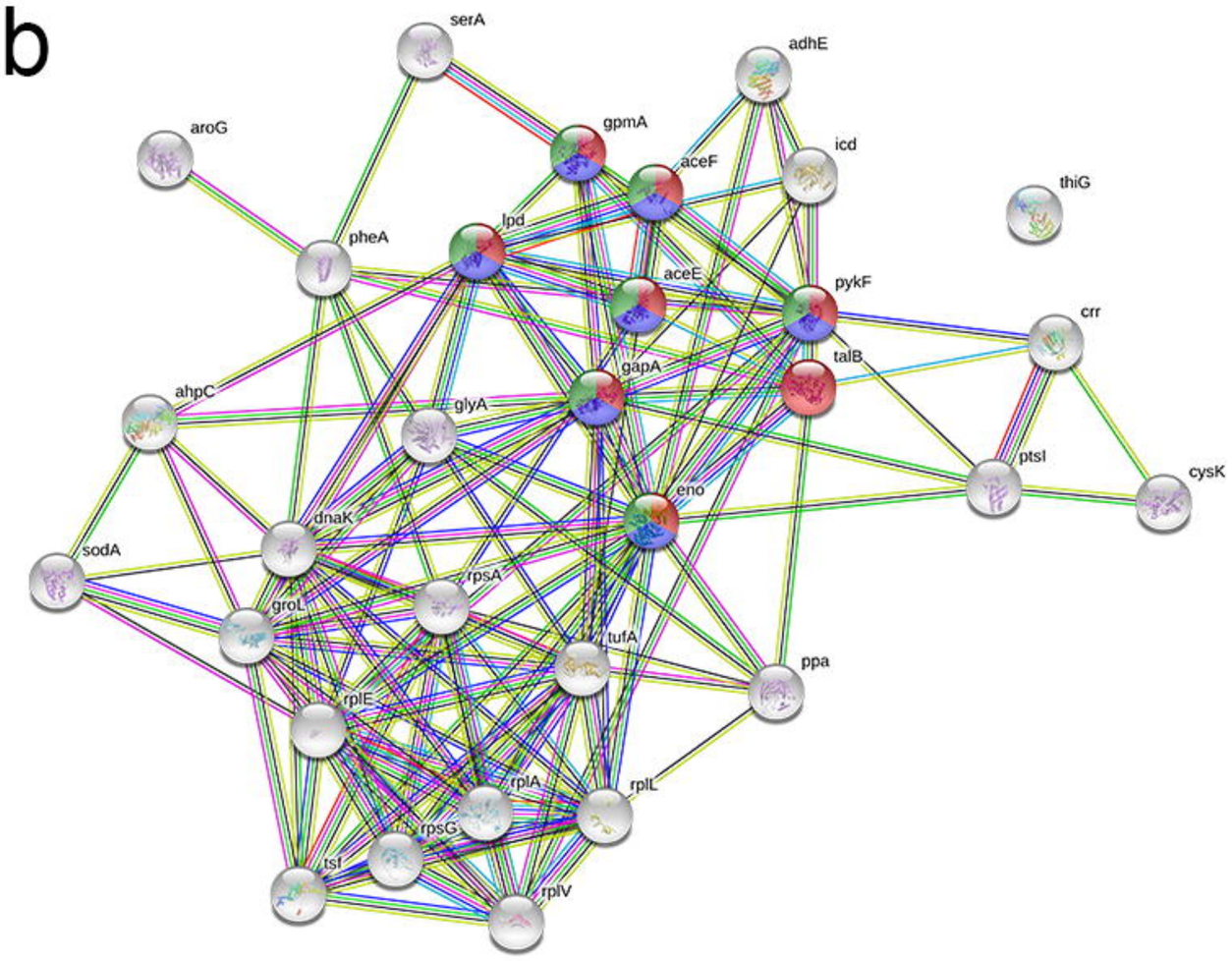
a



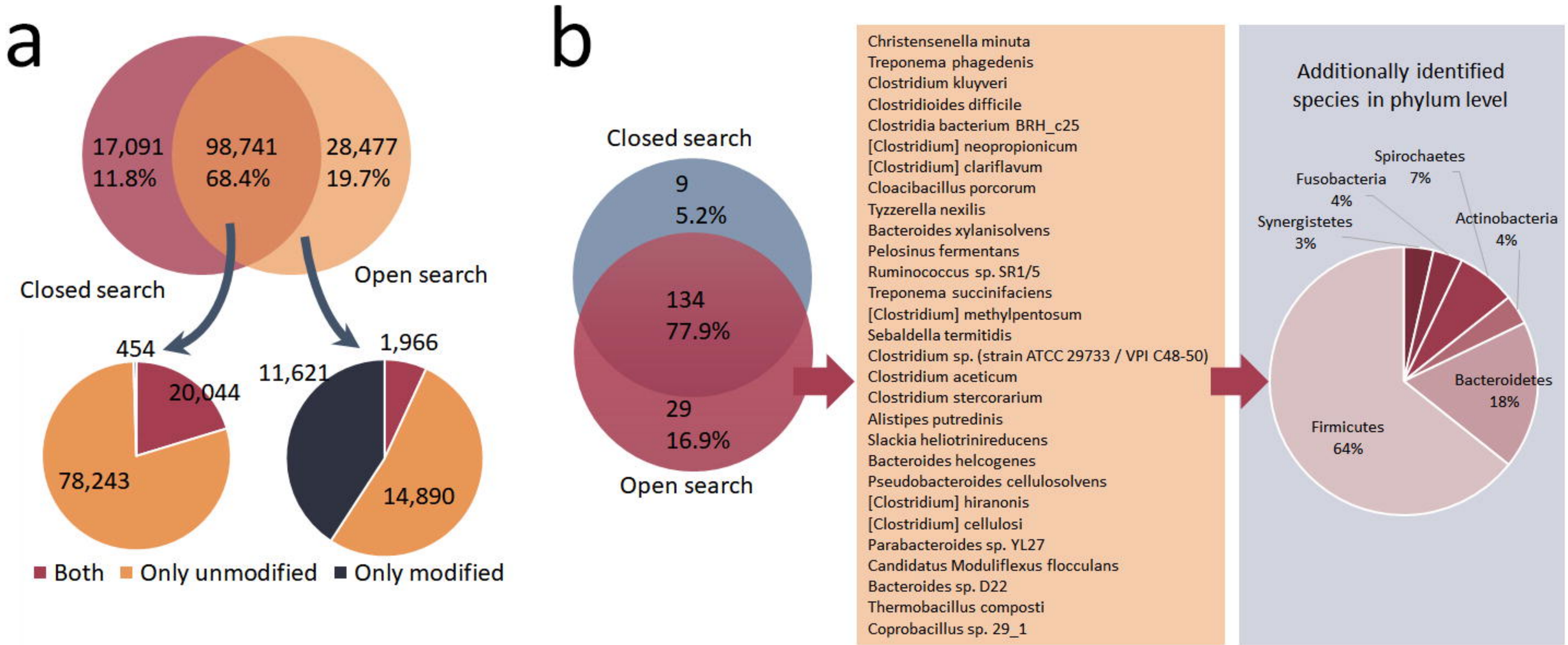
* modification on protein n-term which lose the first amino acid Met, so the actual ΔMass is 20.9481 + 131.0405 = 151.9886.

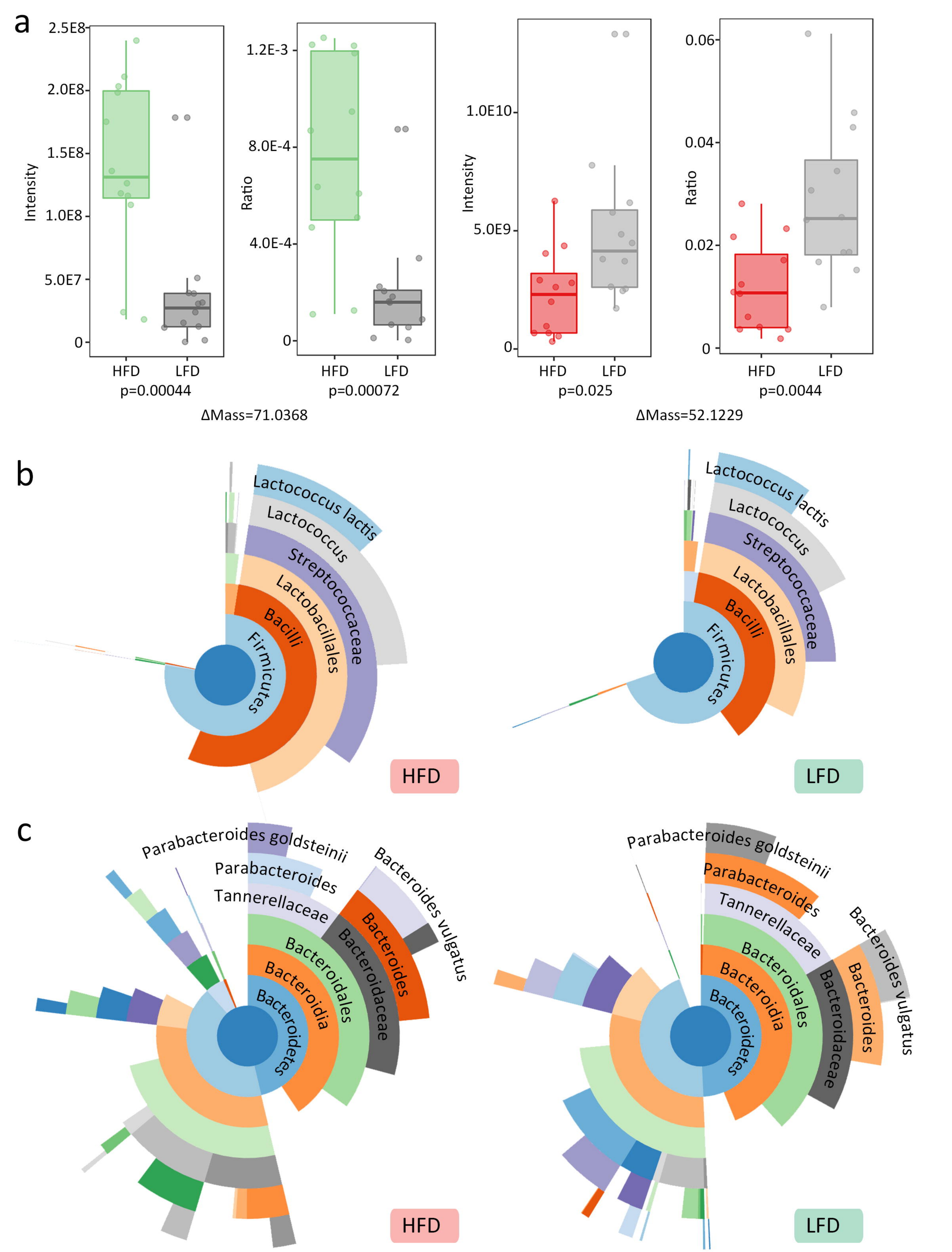
\# modification on Cys (fixed modification site by Carbamidomethyl in database search), so the actual ΔMass is 22.945 + 57.0215 = 79.9665.

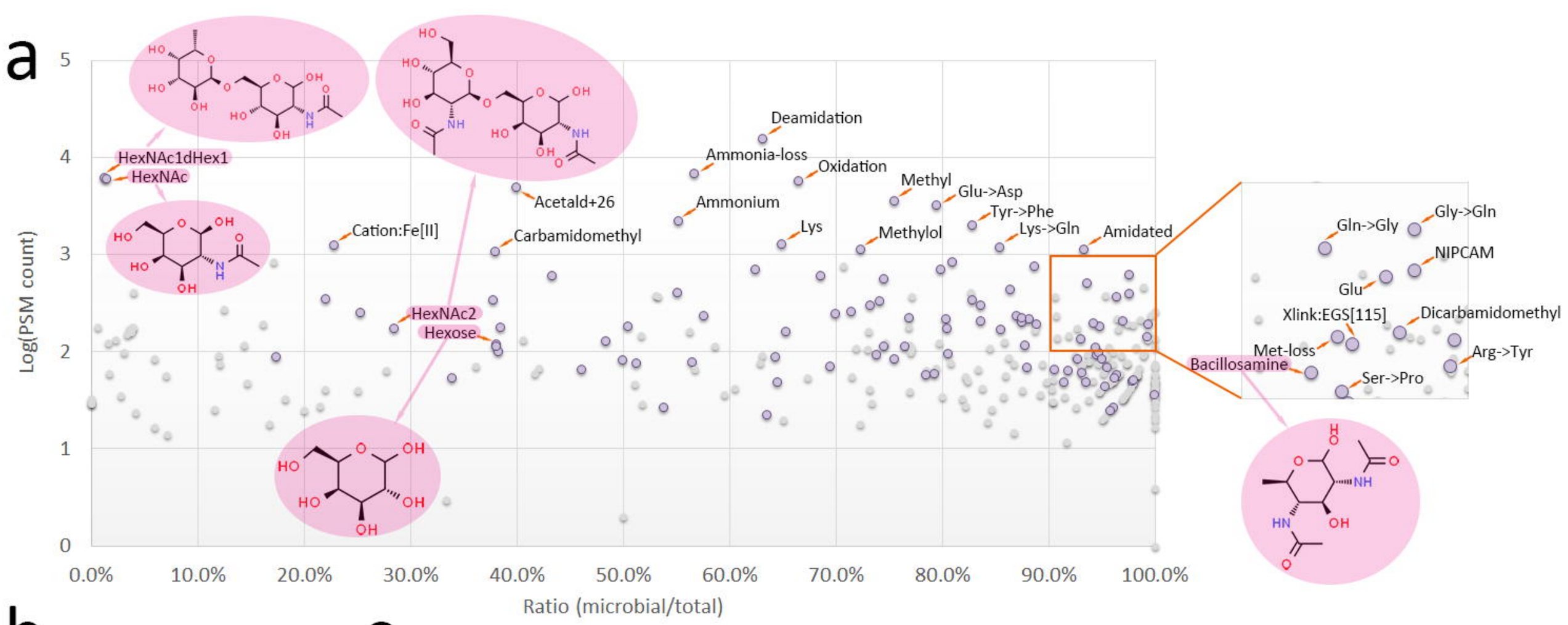
b



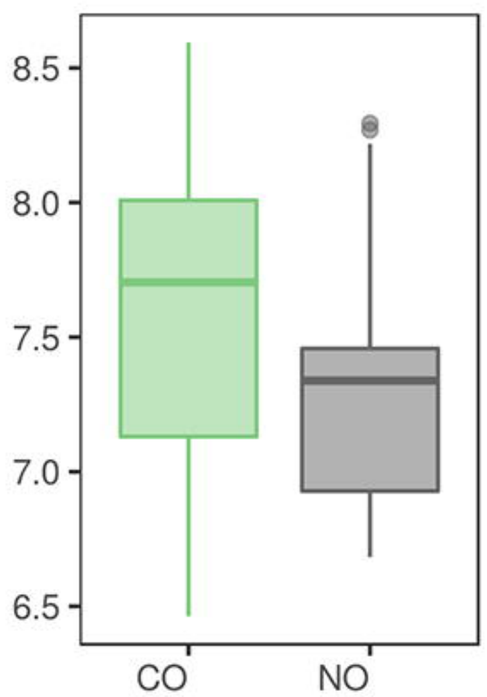
#term ID	term description
GO:1901566	organonitrogen compound biosynthetic process
GO:0006096	glycolytic process
GO:0046496	nicotinamide nucleotide metabolic process
GO:1901564	organonitrogen compound metabolic process
GO:0046394	carboxylic acid biosynthetic process
GO:0044283	small molecule biosynthetic process
GO:0044249	cellular biosynthetic process
GO:1901576	organic substance biosynthetic process
GO:0009168	purine ribonucleoside monophosphate biosynthetic process
GO:0017144	drug metabolic process
GO:0006732	coenzyme metabolic process
GO:0044237	cellular metabolic process
GO:0009987	cellular process
GO:0043603	cellular amide metabolic process
GO:0008152	metabolic process
GO:0006412	translation
GO:0043604	amide biosynthetic process
GO:0019693	ribose phosphate metabolic process
GO:0009108	coenzyme biosynthetic process


a
17,091
11.8%
98,741
68.4%
28,477
19.7%
Closed search
Open search
454
20,044
78,243
11,621
1,966
14,890
Both
Only unmodified
Only modified
b
Closed search
9
5.2%
134
77.9%
29
16.9%
Open search
Christensenella minuta
Treponema phagedenis
Clostridium kluyveri
Clostridioides difficile
Clostridia bacterium BRH_c25
[Clostridium] neopropionicum
[Clostridium] clariflavum
Cloacibacillus porcorum
Tyzzerella nexilis
Bacteroides xylanisolvens
Pelosinus fermentans
Ruminococcus sp. SR1/5
Treponema succinifaciens
[Clostridium] methylpentosum
Sebaldella termitidis
Clostridium sp. (strain ATCC 29733 / VPI C48-50)
Clostridium aceticum
Clostridium stercorarium
Alistipes putredinis
Slackia heliotrinireducens
Bacteroides helcogenes
Pseudobacteroides cellulosolvens
[Clostridium] hiranonis
[Clostridium] cellulosi
Parabacteroides sp. YL27
Candidatus Moduliflexus flocculans
Bacteroides sp. D22
Thermobacillus composti
Coprobacillus sp. 29_1
Additionally identified species in phylum level
Spirochaetes 7%
Fusobacteria 4%
Synergistetes 3%
Actinobacteria 4%
Bacteroidetes 18%
Firmicutes 64%


a
Intensity
2.5E8
2.0E8
1.5E8
1.0E8
5.0E7
0
HFD
LFD
p=0.00044
Ratio
1.2E-3
8.0E-4
4.0E-4
0
HFD
LFD
p=0.00072
ΔMass=71.0368
Intensity
1.0E10
5.0E9
0
HFD
LFD
p=0.025
Ratio
0.06
0.04
0.02
0
HFD
LFD
p=0.0044
ΔMass=52.1229
b
Lactococcus lactis
Lactococcus
Streptococcaceae
Lactobacillales
Bacilli
Firmicutes
HFD
Lactococcus lactis
Lactococcus
Streptococcaceae
Lactobacillales
Bacilli
Firmicutes
LFD
c
Parabacteroides goldsteinii
Parabacteroides
Tannerellaceae
Bacteroides vulgatus
Bacteroides
Bacteroidaceae
Bacteroidales
Bacteroidia
Bacteroidetes
HFD
Parabacteroides goldsteinii
Parabacteroides
Tannerellaceae
Bacteroides vulgatus
Bacteroides
Bacteroidaceae
Bacteroidales
Bacteroidia
Bacteroidetes
LFD



b

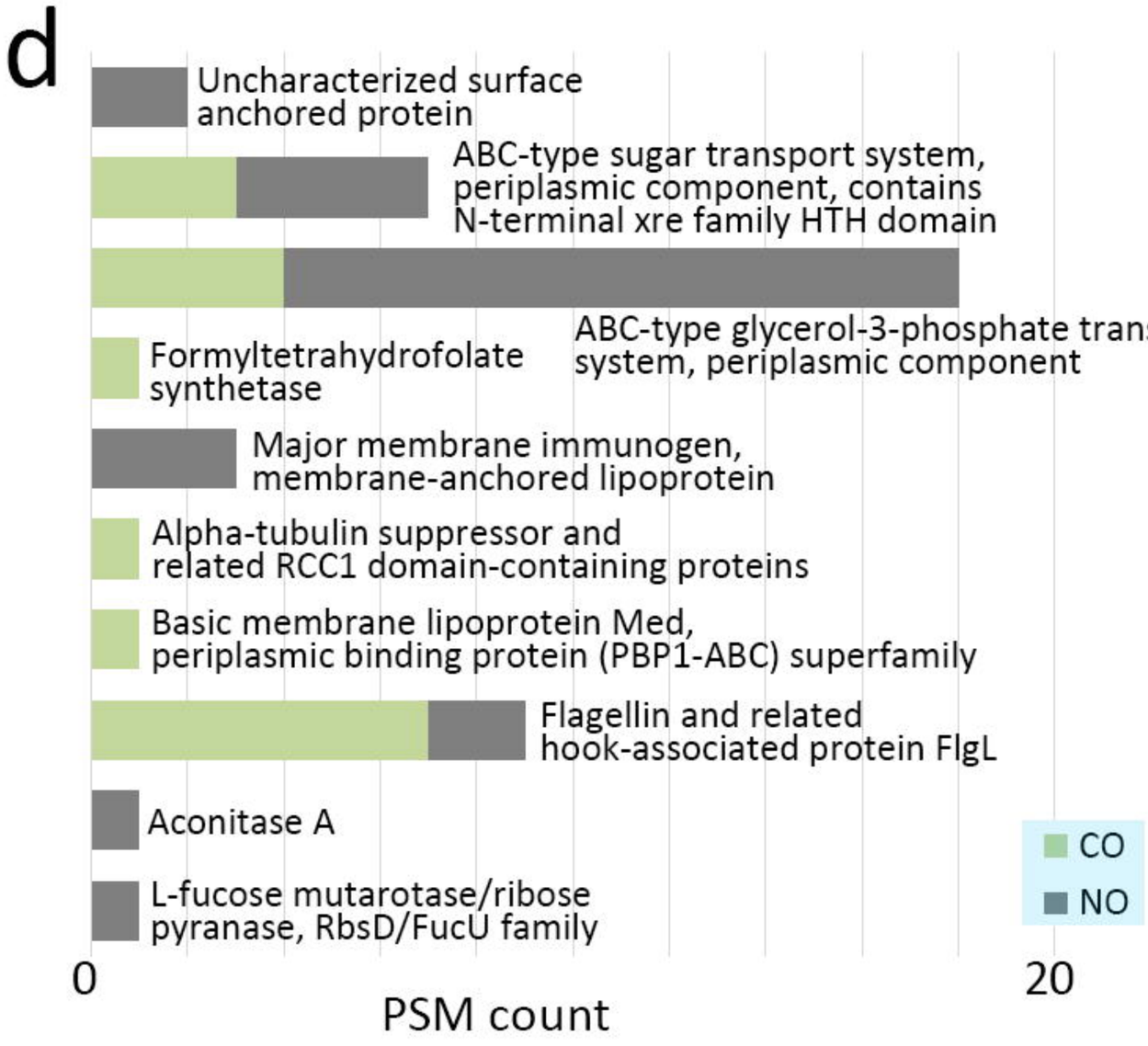
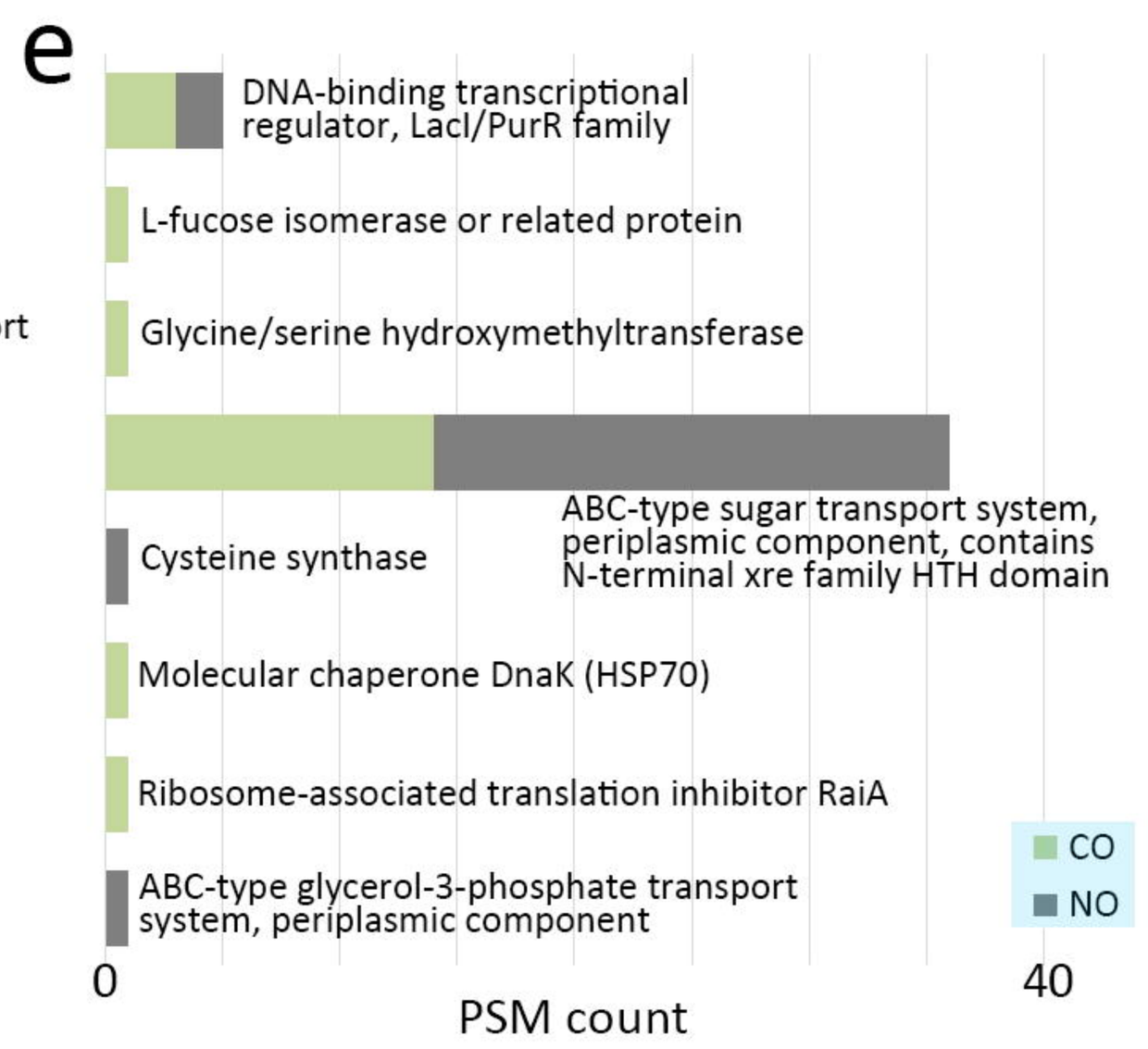


P = 0.048

c

GO-term	Description	Count in gene set	False discovery rate
GO:0002283	neutrophil activation involved in immune response	32 of 489	5.96E-22
GO:0002446	neutrophil mediated immunity	32 of 498	5.96E-22
GO:0043312	neutrophil degranulation	31 of 485	1.06E-21
GO:0002366	leukocyte activation involved in immune response	33 of 616	4.09E-21
GO:0045055	regulated exocytosis	34 of 691	7.65E-21
GO:0045321	leukocyte activation	36 of 894	1.61E-19
GO:0002252	immune effector process	36 of 927	4.48E-19
GO:0046903	secretion	36 of 1070	3.77E-17
GO:0002376	immune system process	50 of 2370	9.78E-17
GO:0006955	immune response	40 of 1560	2.04E-15
GO:0016192	vesicle-mediated transport	40 of 1699	3.45E-14
GO:0006508	proteolysis	34 of 1203	6.35E-14
GO:0050435	amyloid-beta metabolic process	7 of 16	2.13E-09
GO:0035333	Notch receptor processing, ligand-dependent	6 of 7	2.93E-09
GO:0016485	protein processing	12 of 142	3.21E-09
GO:1901564	organonitrogen compound metabolic process	61 of 5281	3.25E-09
GO:0006810	transport	52 of 4130	1.42E-08
GO:0002682	regulation of immune system process	29 of 1391	1.59E-08
GO:0043603	cellular amide metabolic process	21 of 732	3.29E-08
GO:0051179	localization	57 of 5233	2.38E-07

d



e