MARLOWE: Taxonomic Characterization of Unknown Samples for Forensics Using *De Novo* Peptide Identification

Sarah C. Jenson[1†], Fanny Chu[1†*], Anthony S. Barente[1,2], Dustin L. Crockett[3], Natalie C. Lamar[4], Eric D. Merkley[1], Kristin H. Jarman[1‡]

Affiliations:

[1]Chemical & Biological Signatures Group, Pacific Northwest National Laboratory, Richland, WA 99354

[2]Department of Genome Sciences, University of Washington, Seattle, WA 98194

[3]Applied Decisions Systems and Analytics, Group, Pacific Northwest National Laboratory, Richland, WA 99354

[4]Applied Statistics and Computational Modeling Group, Pacific Northwest National Laboratory, Richland, WA 99354

[†]These authors contributed equally to this work.

***Corresponding author:**

Fanny Chu

fanny.chu@pnnl.gov

509.372.4819

[‡]Current affiliation:

Kristin H. Jarman

Karius Inc.

975 Island Dr., Suite 101

Redwood City, CA, 94065

Abstract

We present a computational tool, MARLOWE, for source organism characterization of unknown, forensic biological samples. The intent of MARLOWE is to address a gap in applying proteomics data analysis to forensic applications. MARLOWE produces a list of potential source organisms given confident peptide tags derived from *de novo* peptide sequencing and a statistical approach to assign peptides to organisms in a probabilistic manner, based on a broad sequence database. In this way, the algorithm assumes no *a priori* knowledge of potential sources, and the probabilistic way peptides are taxonomically assigned and then scored enables results to be unbiased (within the constraints of the sequence database). In a proof-of-concept study, we examined MARLOWE's performance on two datasets, the biodiversity dataset and the *Bacillus cereus* superspecies dataset. Not only did MARLOWE demonstrate successful characterization to true contributors in single source and binary mixtures in the biodiversity dataset, but also provided sufficient specificity to distinguish species within a bacterial superspecies group. These results suggest that MARLOWE is suitable for candidate- or lead-generation identification of single-organism and binary samples that can generate forensic leads and aid in selecting appropriate follow-on analyses in a forensic context.

Keywords

Forensic proteomics

Metaproteomics

Strong peptides

De novo sequence tags

Mass spectrometry

Introduction

The last ten years has seen an increase in the use of proteomics methods for forensic applications [1, 2], including sex estimation in human remains using sex-specific forms of amelogenin in tooth enamel [3], identification of individuals by genetically variant peptides in hair proteins [4, 5] or touch samples [6], biofluid identification [7, 8], and protein toxin detection [9, 10]. This paper focuses on identification (taxonomic characterization) of unknown forensic samples using bottom-up proteomics, which is closely related to the identification of microorganisms motivated by clinical needs. It is also related to metaproteomics (proteomics analysis of microbial communities in environmental or clinical samples), but forensics samples are not always complex communities; seized commodities, wildlife remains, or organisms cultivated for illegal purposes may contain only one or a few species.

Forensic samples present special challenges that rarely occur in fundamental science research. If the organism that produced the sample is unknown, the protein sequence database that should be used to identify peptides and proteins in the sample is also unknown, which is effectively the same case as the study of an organism with an unsequenced genome. Analysis of unknown samples may be qualitative, quantitative, or both. Qualitative analysis seeks to identify or at least provide some taxonomic indication of the organism(s) present; quantitative analysis seeks to characterize the abundance (biomass fraction) of the various organisms present. Both types of analysis can be important in various applications, but in many forensic contexts, the emphasis is on qualitative analysis, because laws in various jurisdictions may prohibit specific organisms without reference to quantity. In this paper, we focus primarily on qualitative analysis, and refer to this process as taxonomic characterization. In addition to forensics,

taxonomic characterization is also relevant to clinical, environmental, and archaeological/cultural applications.[11]

Several software tools or algorithms for taxonomic characterization of unknown samples using bottom-up proteomics and a database search have been described, but this general approach requires careful consideration of database composition and shared peptide assignment. Relevant software tools include ABOID [12], TCUP [13], MiCId [14-18], phyloproteomics [19, 20], a statistical approach described by Jarman et al. [21], and two approaches specific for species identification from bone samples by Rüther et al. [22] and Yang et al. [23]. All these protocols focus on analysis of untargeted, data-dependent, bottom-up proteomics data using database search with either a broad, multispecies database (either as a single database, or as multiple parallel searches with individual organism databases). The database may focus on particular organisms (e.g., pathogens) depending on the application. After the database search, peptide-spectrum matches (PSMs) or peptide sequences are assigned to organisms, and potential organisms are scored based on that mapping. However, the variations on this general framework all face two problems: database creation, and how to account for peptides that are shared between organisms.

Database Creation. There are known issues with searching a very large sequence database.[24] As the number of organisms' proteomes included in the database grows, the search space increases, leading to a higher number of false positive PSMs based purely on statistics. To control for false positive PSMs via the false discovery rate, score thresholds are increased, which leads to decreased sensitivity. (It is worth mentioning that MiCId [25] uses a different statistical approach meant to address this issue.) Conversely, if the database is too small, the correct organism may not be present, leading to a failure to identify it. Worse, an unrealistically narrow database can lead to an incorrect or unrealistically specific identification, such as apparently

identifying a strain when only species is realistic. [26, 27] All of the general tools mentioned here use a broad or even comprehensive database initially, such as all available microbial genomes (e.g., ABOID,[12]) or the NCBI non-redundant database. Phylopeptidomics starts with a broad database, then conducts a second-round search using only organisms detected in the first round. [19] The SPIN method [22], designed for identifying mammals via bone fragment proteomics, constructs a multispecies database of a small number of proteins found in mammalian bones. Clearly, there is no standardized approach for database creation; bespoke databases are created to suit each application and tool, but their composition needs to be carefully considered to avoid the issues with false positives and false negatives described above.

Shared peptides. The second issue centers on assignment of shared peptides, that is, peptides that appear in the protein sequences of multiple organisms. Peptides that are unique to a single taxon in the database necessarily indicate that taxon, or one that is not present in the database. The simplest approach is to use only unique peptides, but these may not exist, or may not be expressed or detected; this approach can severely limit the ability to identify source organisms. On the other hand, many more peptides are shared among organisms than are unique to a single organism. Can shared peptides instead augment the presence of unique peptides and help identify taxa, given that peptides that appear in multiple taxa tend to occur in related taxa?

Several algorithms and tools exploit information from shared peptides to inform taxa, in different ways. TCUP [13] and UniPept [28] use the lowest common ancestor algorithm, which assigns peptides to the most basal node of the taxonomic tree to which they are unique. After matching PSMs to a MySQL database, ABOID uses a combination of clustering and principal components analysis to determine the most likely organism. [12]. MiCId [14] also clusters organisms that share peptides, and weights the calculated p-value for each peptide based on an inverse

function of the number of clusters in which that peptide appears. Recent efforts to augment organism identification functionality in MiCId have focused on algorithm improvements [15, 16], usability [17], and extensibility to different types of proteomics data [18]. Phylopeptidomics, using an iterative database creation and search approach that increases in specificity per iteration, leverages a mapping of mass spectra to taxa, called taxon-spectrum matches (TSMs), which is analogous to peptide-spectrum matches (PSMs). This approach has been used to identify pathogens in archeological bone samples [29] and to characterize the gut microbiome of COVID-19 patients [19]. This same research group has also expanded phylopeptidomics to estimate the biomass of each species [20] by plotting TSMs for a hypothesized species as a function of the phylogenetic distance from a larger group of species and fitting a multiexponential decay function to the resulting curve. This is a unique way to derive useful taxonomic information from shared peptides and the amplitude parameters from this fit reflect the known biomass ratios. Finally, Jarman et al. [21] have created an approach that introduced strong peptides, that is, peptides that are not uniquely assigned to a single taxon, but that are strongly biased to appear in only a few taxa. Phylogenetic organism clusters produced using the shared and strong peptides concept approximates "supergenus" or "superspecies" groups, as an alternative to classical taxonomy (e.g., genus, species). Again, as with database creation, there is no standardized approach to leveraging shared peptide assignments to identify organisms, but all recognize the value of shared peptides to inform taxonomy from proteomics data.

In this paper, we introduce an approach that provides unique answers to these two questions. First, we attempt to decrease the reliance on sequence databases by deploying *de novo* peptide identification [30-32]. Because *de novo* methods often lead to only partially correct peptides, we use only confident sequence regions (tags) to match to potential organisms. This

approach separates at least the peptide identification step from the limitations of existing databases, although matching sequence tags to organisms still requires a database of peptide sequences. Several researchers in metaproteomics have explored the use of *de novo* peptide identifications in various ways [33-35]; the current paper attempts to use *de novo* sequencing for unknown identification in a forensic context. We combine *de novo* sequencing with (1) an adaptation of the strong peptide concept and (2) a novel statistical procedure for assigning peptides to organisms that accounts for shared peptides. Here, our algorithm estimates the number of peptides detected in a sample that could belong to an organism or its nearest neighbors based on proteomic similarity even if that particular organism were absent from our database.

Briefly, our tool, which we call MARLOWE (after the fictional private investigator created by author Raymond Chandler, in honor of the intended investigative use of the tool), uses *de novo* peptide identification, tag extraction and filtering by peptide strength, assignment of tags to taxonomic sources, adjustment of tag counts for tags shared between organisms, and final scoring of source organisms using non-negative least squares regression. Rather than relying solely on unique peptides, this procedure appropriately leverages the information contained in shared peptides by using the peptide strength concept referenced above.

In the proof-of-concept examination of MARLOWE's performance presented here, we find that MARLOWE correctly characterizes most contributors of microbial samples in single-source and binary mixtures at the species level. Further, it is also able to correctly characterize microbial species within the closely related bacterial group *Bacillus cereus* to the species level, demonstrating its specificity. These results demonstrate that MARLOWE is capable of

preliminary, qualitative organism identification/taxonomic characterization of unknown protein-containing samples. As such, it fills an important gap in forensic proteomics.

Methods

The MARLOWE Algorithm

MARLOWE uses the following steps: (1) *de novo* peptide identification, (2) tag extraction and peptide strength filtering, (3) tag matching to a broad sequence database, (4) tag assignment to organisms, including a correction for the number of tags expected due to shared peptides if the organism is not present, and (5) taxonomic group scoring. Each of these steps is described below, along with a detailed explanation of the concept of peptide strength.

De novo peptide identification. With minor output file formatting, MARLOWE can accept input from PEAKS, [36] Novor,[37] or Casanovo. [38] The only requirement is that the *de novo* tool provide (in addition to the usual PSM data) a local or per-amino acid residue confidence score that reflects the likelihood that an individual residue is correct. The listed tools all provide this local confidence score. In this paper, we used Novor.

Tag extraction. The accuracy of *de novo* PSMs is known to be lower than database search PSMs. However, *de novo* PSMs often contain correct subsequences [39] called tags, which have been used in various peptide identification schemes for a long time [40, 41]. In MARLOWE, we define a tag as a stretch of at least 5 consecutive amino acid residues in which each residue has a local confidence score of 80 (out of 100) or greater. Only tags from peptides where the overall PSM had an average local confidence score of 50 or greater were used; this ensured that only high-quality spectra were considered.

Peptide strength filtering. To optimize the taxonomic specificity available from *de novo* sequence tags, we calculate peptide strength for each and then filter them to retain only tags that

match to peptides (strong tags). To understand the concept of strong peptides and their role in MARLOWE, it is important to keep the following points in mind:

1. Peptide strength is a way to avoid the limitations of "unique" peptides. Both genomics [42] and proteomics studies [43, 44] have demonstrated that, as databases grow, sequences used as biomarkers that were once unique lose that uniqueness.
2. Peptide strength is dependent on the database being used, but far less dependent than peptide uniqueness. Jarman et al. showed that when new, related organisms are added to a sequence database, the number of unique peptides declines precipitously, but the peptide strength remains nearly constant. Thus, while peptide strength will change as databases are updated, we do not expect it to change dramatically, and peptide strength calculations from older databases will remain taxonomically informative for a long time.
3. Peptide strength is a way to leverage the information content of non-uniformly shared peptides. Strong peptides are shared peptides, but they are shared in a way that is biased across the taxonomic tree at a given level of taxonomy. Whereas unique peptides are specific to (for instance) a particular species, strong peptides are selective with respect to a group of related species. An accumulation of strong peptides can thus provide a very strong indication of the presence of one or more members of that group. This is very similar to the insight of Armengaud and coworkers [20] that shared peptides are informative in taxonomic characterization, and to the use of weighting factors assigned to shared peptides by Alves et al. in MiCId [14] (although MARLOWE uses a filtering rather than a weighting approach).

The peptide strength (ps) of a peptide sequence ρ is defined by Jarman et al. [21] as:

$$ps(\rho) = \frac{N-1}{N+M-2} \quad (1)$$

where N is the total number of taxa (at a given level) in the database, and M is the number of taxa in the database that contain peptide ρ. This equation is derived using the definitions of likelihood ratio and weight of evidence (statistical concepts often used in forensics), and expressions for the probability that peptide ρ occurs in organisms other than the one currently in question. Interested readers should consult the derivation given in the supplementary material of Jarman et al. [21] Peptide strength varies from ½ (for the case where N = M, meaning that peptide ρ is found in every taxon in the entire database) to unity (for the case where ρ occurs in only a single taxon, i.e., M = 1). Let Mc represent the critical number of taxa in which ρ appears. By setting a threshold value of 0.95 for the peptide strength and assuming N>>1, it follows from Eqn (1) that Mc = 0.053N, meaning that strong peptides occur in no more than 5.3% of the taxa at a given level in the database (5.3% is roughly the ratio 0.05/0.95). Importantly, as more organisms are sequenced and the database grows, as long as ρ appears in fewer than 5.3% of all taxa, it will remain a strong peptide.

This definition of peptide strength worked well for taxonomic characterization using database searching, but we found that using it with a broad sequence database resulted in nearly every peptide being classified as a strong peptide. To achieve more stringent filtering, we further defined a *pairwise* peptide strength as follows. (This description counts species in genera, but an analogous definition can be used for genera within a family or any other similar combination of adjacent taxonomic levels.) For each genus that contains ρ, the number of species in that genus that contain ρ is determined (analogous to M in Equation 1) and divided by the total number of species in that genus (analogous to N). We call this ratio a hit fraction. The genera with the largest and second-largest hit fractions are then identified. If the ratio of the largest and second-

largest hit fractions is greater than 1/0.053, then the peptide is defined as strong. This procedure is equivalent to stating that a peptide is strong if it exists in 95% or more of the species in the genus where it occurs most often, and 5% or fewer of the species in the genus where it occurs the second most often. The pairwise peptide strength concept can be applied to tags as well.

Tag matching to a broad sequence database. To interpret *de novo* peptide identifications as evidence for the presence of particular taxa, *de novo* tags must be matched to a broad sequence database. In this proof-of-concept study, we have employed the KEGG Genomes database (release 91.0, https://www.genome.jp/kegg/genome/, downloaded July 1, 2019). All protein sequences (viral genomes were removed because the small size of viral proteomes would impact taxonomic clustering—see below) from this source were *in silico*-digested with trypsin and stored in a MySQL database. Tags extracted from Novor *de novo* sequencing output were used to query this database and retrieve peptide, protein, organism, and taxonomy tree (lineage) information. The database was optimized to speed execution of these queries.

Tag assignment to taxonomic groups. To avoid the confusion created by the non-phylogenetic nature of formal taxonomy, we group all species in the KEGG database into phylogenetic taxonomic groups based on shared tryptic peptides, which we refer to as taxonomic groups. To accomplish this grouping, we first predict all tryptic peptides for each organism, then create a symmetric matrix (the overlap matrix) consisting of the Jaccard coefficients (overlap coefficients) for the sets of tryptic peptides of each pair of organisms. Taxonomic groups are then formed by clustering using the leader algorithm and Euclidean distance based on the overlap coefficients, with a radius of 0.5.

Tags extracted from *de novo* peptide identification results are first matched to the sequence database to identify those that could belong to genus-level strong peptides. The tag list is then further filtered for tags that match to proteins with at least two tags. Next, the list of tags is filtered for strong tags (not peptides) with respect to the clustered taxonomic groups. This final list of tags is the input for the assignment to taxonomic groups/organisms.

Outside of the peptide strength calculation and after filtering, we assign peptide tags to organisms in the following manner. Specifically, let N_i be the actual (true) number of peptides contributed by organism i in an unknown sample. The expected number of observed counts for organism i, e_i, can be approximated as

$$e_i = \sum_j N_j \, P\{\rho \in O_i | \rho \in O_j\}$$

where $P\{\rho \in O_i | \rho \in O_j\}$ is the probability a randomly selected peptide ρ belongs to organism O_i, given it also belongs to organism O_j. In matrix notation, the expected number of counts for all organisms $\underline{E}$ can be written as

$$\underline{E} = \boldsymbol{P}\underline{N} \tag{2}$$

where $\underline{N}$ is an M x 1 vector containing the actual number of peptides contributed by organism i=1,2,…M, and $\boldsymbol{P}$ is an MxM matrix where entry i,j is $P\{\rho \in O_i | \rho \in O_j\}$, calculated as the number of peptides shared by organisms i and j, divided by the number of peptides belonging to organism j. By substituting $\underline{E}$ with the observed number of counts $\underline{\hat{E}}$, we estimate the actual number of peptides contributed by each organism ($\underline{\hat{N}}$) by inverting Eqn (2),

$$\boldsymbol{P}^{-1}\underline{\hat{E}} = \underline{\hat{N}}$$

Taxonomic group scoring. The above scoring process adjusts the peptide counts associated with each source to reduce confusion caused by phylogenetic similarity between

organisms (i.e., co-identifications). After tags are assigned to taxonomic groups, these tabulated assignments and an organism-level overlap matrix are used to assign scores via the non-negative least squares algorithm. In this step, the count of matching tags is corrected to the final taxonomic score based on the overlap coefficients. For any two organisms *i* and *j*, with *i* present in the sample and *j* absent, the expected number of hits for organism *j* is the product of the number of hits for *i* and the overlap coefficient. If fewer hits are found, the score is reduced. If more hits are found, the score is increased. MARLOWE reports both the corrected hit counts (taxonomic) scores for each detected taxonomic group, and, as a secondary score, the raw hit counts per species. By applying these statistical concepts to tryptic peptide tags from the KEGG database, Release 91.0 (July 1, 2019), MARLOWE narrows down the protein sources in a sample, enabling a more focused follow-on confirmatory analysis.

Publicly available raw mass spectrometry data

Raw spectrometry data files from the project PXD003669 were downloaded from ProteomeXchange and can be accessed via the ProteomeXchange online repository[45, 46] (http://proteomecentral.proteomexchange.org). Data from the PNNL Biodiversity Library[47] can be accessed from ProteomeXchange via the MassIVE repository with identifiers PXD001860 and MSV000079053, although we accessed this data using an internal PNNL system. These datasets were converted to MGF format using ThermoRawFileParser.

MARLOWE implementation and execution time

MARLOWE is implemented in R (version 4.2.2) via three R packages: MakeSearchSim, CandidateSearchDatabase, and OrgIDPipeline. MakeSearchSim contains the functionality necessary to run Novor, the *de novo* peptide sequencing tool, and parse its output. CandidateSearchDatabase contains all functionality relating to the creation and use of the

MySQL database that houses a processed version of the KEGG database (Release 91.0 accessed July 2019). OrgIDPipeline contains the functions related to implementation of the statistical algorithms and methods necessary to (1) calculate the final taxonomic scores from the database query results and (2) generate the final list of potential source organisms.

An important consideration in algorithm usage is execution time. Here, we describe time estimates for use of MARLOWE itself, which is distinct from the time needed for *de novo* peptide sequencing, also described. This is because various *de novo* peptide sequencing tools exist and as such, there was no need for us to develop a bespoke tool specifically to integrate into MARLOWE for the same purpose. Further, MARLOWE provides user flexibility in its ability to utilize input from a few different *de novo* sequencing tools. For a mass spectrometry dataset, consisting of a collection of raw data files, MARLOWE requires on average, 4 minutes per data file, from ingesting *de novo* peptide sequence lists to producing a list of potential source organisms. Execution time for each individual third-party de novo sequencing tool will vary. In our hands, we have observed an average of 6 minutes per raw data file for Novor, and an execution time of up to 30 minutes per raw data file using the commercial tool PEAKS Online.

Results & Discussion

Proof-of-Concept: Biodiversity

We demonstrate the application of MARLOWE in a diverse set of 66 pure bacterial cultures and 25 simulated binary mixtures each of varying ratios (i.e., tandem mass spectra from 20, 30, or 40 proteins belonging to secondary source appended to primary source). Each simulated dataset was generated using a different random sampling of 20, 30, or 40 contributor proteins. Results are listed in Table 1 below.

Table 1. Summary of MARLOWE's performance for single-source (pure) and simulated binary mixtures.

Dataset	Number of Samples	Primary Contributor Characterization Rate (% top 1/top2/top5)*			Secondary Contributor Characterization Rate (% top2/top5)†		
		Genus	Species	Taxonomic Group	Genus	Species	Taxonomic Group
Biodiversity Pure	66	100/-/-	94/100/-	100/-/-	--	--	--
Biodiversity Simulated 20	25	100/-/-	76/100/-	100/-/-	48/72	20/60	48/72
Biodiversity Simulated 30	25	100/-/-	76/100/-	100/-/-	56/76	24/68	56/76
Biodiversity Simulated 40	25	100/-/-	76/100/-	100/-/-	60/76	28/68	60/76

* Correct characterization is based on hit to true contributor as the top N-ranked result for taxa included in KEGG

† Correct characterization rates for contributor organisms, or secondary sources, are determined by representation of the true secondary source organisms within the top 2/top 5 hits, respectively.

For pure cultures, we achieved 100% correct characterization specific to the species level; correct characterization is defined as MARLOWE returning the correct taxonomy in a ranked organism list (top organism, top two organisms, or top five organisms as show in Table 1). Of these correct characterizations, 100% was achieved with the correct genus and the correct taxonomic group as the top scoring organism, and 94% achieved with the correct species as the top scoring organism. 100% of the correct species were represented in the top 2 lists of highest scoring organisms.

In simulated mixtures, the primary source was 100% correctly identified at the genus and taxonomic group levels as the top-ranked hit for the various mixture ratios. At the species level, 100% correct characterization of the primary organism at the species level occurred when considering the top 2 hits. For the secondary sources, correctly identified secondary sources

within the top 5 hits range between 72% and 76% at the genus and taxonomic group levels, and between 60% and 68% at the species level, across the various mixture ratios. Both the primary and contributor organisms in mixtures can be correctly characterized at a maximum rate of 76%. Characterization of mixtures is more successful when there is a larger contribution from secondary sources and when considering at taxonomic group or genus level compared to species-level. These results demonstrate initial implementation of the above statistical concepts in MARLOWE for successful organism source characterization.

Characterization Specificity: B. cereus group

We then examined MARLOWE's performance—in particular, specificity—on a closely related bacterial group (Table 2). Characterization specificity was assessed using a dataset comprising seven species within the *Bacillus cereus* group that was described in Pfrunder et al. (2016).[44] Strains include *B. cereus*, *B. cytotoxicus*, *B. mycoides*, *B. pseudomycoides*, *B. thuringiensis*, *B. anthracis* (not analyzed), and *B. toyonensis*, which all share great genomic similarity.[48] These strains span a range of pathogen toxicity and health safety concerns,[49] and as such, the ability to distinguish and correctly classify unknown samples to the different species in the group is of forensic, biosecurity, and health interest. Pfrunder and coworkers also examined *B. weihenstephanensis*, a strain within the *B. cereus* group that has been taxonomically reassigned to fall under *B. mycoides* species.[50] In this analysis, the true contributor of *B. weihenstephanensis* samples at the species level is therefore assigned as *B. mycoides*. Characterization at the species level was chosen as a sufficiently specific taxonomic level since strains of the same species usually exhibit high degrees of genetic similarity. Further, actual strains of true contributors may not be included in the KEGG Genome database that forms the basis for MARLOWE, such as in the case of *B. weihenstephanensis*.

Table 2. Characterization summary of MARLOWE for *B. cereus* superspecies group with known true contributors

Dataset	Number of Samples	Characterization Rate (% top 1/top2/top5)*				Dataset Reference
		Family	Genus	Species	Taxonomic Group	
B. cereus	42	90/93/95	90/93/95	88/90/93	88/90/93	Pfrunder et al. (2016)[44]

* Correct characterization is based on hit to true contributor as the top N-ranked result for taxa included in KEGG

To examine MARLOWE's ability to specifically characterize the true contributor, characterization rate to the true contributor and taxonomic scores to the true contributor compared to all other taxa were quantified. We find that most samples yielded correct characterization to the true contributor *B. cereus* group member with high taxonomic scores, though characterization to *B. cereus* posed a challenge. Comparison of ranked organism lists among these samples showed high proteome diversity among the *B. cereus* group members, which are organized into different and distinct taxonomic groups. MARLOWE displays high characterization specificity for most samples containing *B. cereus* species, as lists of potential sources often unambiguously point to one *B. cereus* group member with a high taxonomic score, indicating great confidence in the characterization.

The true contributor at the species level was correctly characterized as the top-ranked organism in 88% of samples (Figure 1). Specifically, all biological and technical replicates of samples from *B. cytotoxicus*, *B. mycoides*, *B. weihenstephanensis*, *B. pseudomycoides*, and *B. thuringiensis* were correctly characterized to their respective true contributor as the organism with the highest taxonomic score and greatest number of tag-strong peptide matches. Note that samples containing *B. weihenstephanensis*, a strain reassigned as a subspecies of *B. mycoides*,

were correctly characterized as *B. mycoides*. Interestingly, these species with correct characterization to the top-ranked organism had substantially high taxonomic scores (on average, 0.68 ± 0.24 (s.d.) normalized score) compared to all other ranked taxa. As discussed in the previous section, high taxonomic scores indicate greater confidence of the ranked organisms as true contributors, and the taxonomic score distributions for these samples suggest a single contributor sample rather than a mixture.

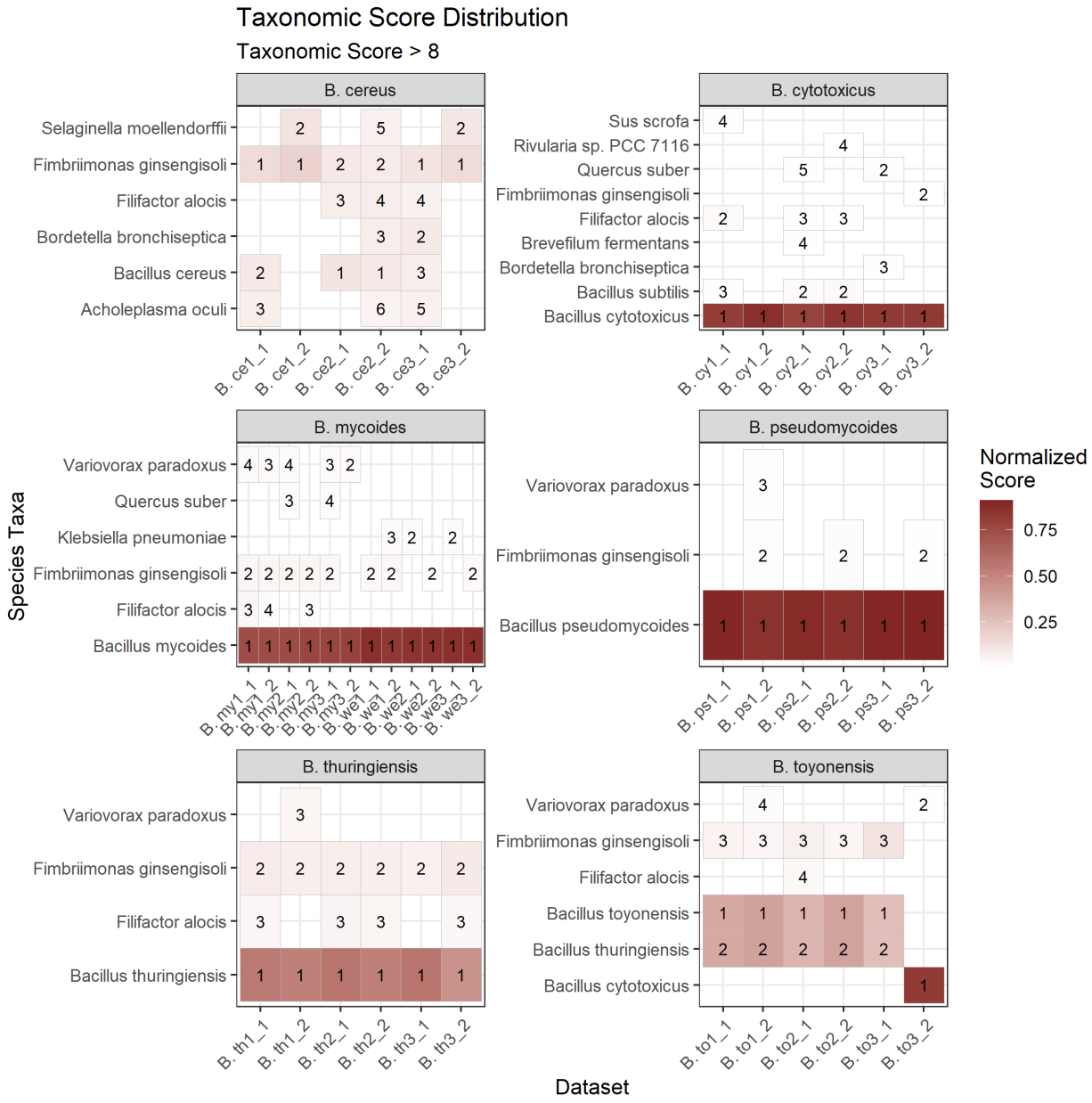

Figure 1. Heatmap of organism score ranking of all taxa with raw taxonomic scores > 8 for each sample within the *B. cereus* group, further grouped by the true contributor at the species level. Heatmap colors are based on normalized taxonomic score, normalized to the total taxonomic score per sample. Values in each cell represent the rank of the organism, assigned using taxonomic score and tag matches to strong peptides. MARLOWE achieved 88% correct characterization of the true contributor as the top-ranked organism in this dataset at the species level. For most samples, taxonomic scores of the true contributor are substantially higher than for all other taxa, indicating confident characterization.

For groups of samples where the true contributor was not the top-ranked organism, among the *B. toyonensis* samples, the characterization rate was 83%, as one of six replicates was characterized as *B. cytotoxicus* with a high taxonomic score (normalized score = 0.82). The high taxonomic score to *B. cytotoxicus* returned for one of the six replicates of *B. toyonensis* suggests an error in sample preparation rather than mischaracterization by MARLOWE; the distribution of taxonomic scores for this sample is also vastly different from the other *B. toyonensis* samples. A substantially greater number of peptide-spectrum matches to *B. cytotoxicus*-specific proteins (3935) compared to *B. toyonensis*-specific proteins (30) from conventional database searching confirms the sample preparation error for that replicate.

Characterization of *B. cereus* samples posed a challenge for MARLOWE, as only two of six replicates ranked the true contributor as the top organism, with *B. cereus* correctly characterized within the top three organisms in four of the six replicates. The results for *B. cereus* samples are also distinct in that the taxonomic scores for this set of samples are generally low and with a minimum taxonomic score of 6.3 (on average, 0.09 ± 0.02 (s.d.) normalized score and 11.4 ± 4.1 (s.d.) taxonomic score). The low taxonomic scores for *B. cereus* samples are due

to the structure of the database and on taxonomic classifications. There are many strains of *B. cereus* in MARLOWE's database (i.e., 14 strains of *B. cereus* split among 5 taxonomic groups). The presence of so many similar peptides in different species results in fewer strong peptides assigned to each specific strain of *B. cereus*, and thus similarly lower taxonomic scores overall. This illustrates the subtle dependence of MARLOWE results on database contents and the similarity of organisms across taxa. Possibly, this score could be improved by using groups constructed by phylogenetic clustering rather than taxonomy assignments. Despite not being the top-ranked organism, *B. cereus* is still returned as a potential organism by MARLOWE within the top three organisms in at least four of six replicates across *B. cereus* samples.

MARLOWE's characterization specificity to members of the *B. cereus* group was examined by comparing taxonomic scores of the true contributor to all other taxa, as well as noting the different species that were ranked. Most notably, except for *B. toyonensis* samples, all ranked taxa other than the true contributor are not members of the *B. cereus* group, and in fact most do not belong to the *Bacillus* genus. Hits to other members of the *B. cereus* group yielded extremely low taxonomic scores (< 1.8 taxonomic score, n = 43 hits across 27 samples), which were filtered out, as the low magnitude of these scores represents spurious tag-peptide matches to peptides not distinctly strong to a single taxonomic group and indicates low confidence of the organism as a contributor in the sample. This suggests that proteomes of each species, particularly the strong peptide distributions, within the *B. cereus* group are sufficiently diverse from other *B. cereus* members and are also distinct from the proteomes of organisms in the *Bacillus* genus. As for the *B. toyonensis* samples, both *B. toyonensis* and *B. thuringiensis* were returned from MARLOWE for the samples that were correctly characterized. Further analysis of their taxonomic scores revealed that the two species belong to the same taxonomic group, and as

such, have identical taxonomic scores. *B. toyonensis* was ranked higher than *B. thuringiensis* owing to a greater number of tag-strong peptide matches (on average, 75 ± 45 (s.d.) matches for *B. toyonensis* compared to 64 ± 39 (s.d.) matches for *B. thuringiensis*). In fact, *B. toyonensis* and *B. thuringiensis* MC28 reside in the same taxonomic group, but a different strain of *B. thuringiensis*, *B. thuringiensis* serovar chinensis CT-43, belongs to a different taxonomic group. The latter strain is the top-ranked organism for *B. thuringiensis* samples. Comparison of their genome sequences via NCBI BLAST showed 99% genetic similarity between *B. toyonensis* and *B. thuringiensis* MC28, but 97% similarity between the two strains of *B. thuringiensis*. This observation indicates not only the ability for MARLOWE to correctly characterize contributors to the species level, but also that similarity of organism proteomes may diverge from their assigned taxonomy and MARLOWE captures these differences.

MARLOWE use and follow-up actions

The intent of MARLOWE is to provide investigative leads on potential sources given an unknown bioforensic sample, without presumption of potential sources either during reassembly of peptide sequences from mass spectrometry data or during assignment of peptide tags to organisms and using a statistically robust method to do so. Results from the proof-of-concept analysis using the biodiversity dataset as well as the *B. cereus* group dataset demonstrate MARLOWE's ability to correctly characterize major contributors in single species and binary species, with a high degree of specificity. MARLOWE's strengths are in narrowing down to potential source organisms, and as such, would be most suitable at the beginning of a bioinformatics pipeline or investigation. We expect and encourage follow-up actions on the results provided by MARLOWE to include confirmatory analyses, such as via database searching or applying bespoke data analysis pipelines for other qualitative and quantitative

analyses. We believe that MARLOWE can be broadly applicable for several applications in mass spectrometry-based protein analysis, and its modular nature is amenable to being co-opted into other pipelines for bespoke bioinformatics analyses.

Acknowledgments

The authors acknowledge the Environmental and Molecular Sciences Laboratory for liquid chromatography-tandem mass spectrometry data acquisition of samples. Development of MARLOWE's workflow was supported by Department of Homeland Security, Science and Technology Directorate - Homeland Security Advanced Research Projects Agency - Chemical and Biological Division under Contract #HSHQPM16X00216.

Competing Interests

The authors declare no competing interests.

Data Availability

Publicly available raw mass spectrometry data were acquired from ProteomeXchange and can be accessed via the online repository[45, 46] (http://proteomecentral.proteomexchange.org) under accessions PXD001860 and PXD003669.

Author contributions

SCJ: algorithm design and development, data analysis

FC: data analysis, manuscript writing

ASB: algorithm development and testing

DLC: algorithm design

NCH: algorithm development

EDM: algorithm conception and design, manuscript writing

KHJ: algorithm conception, design, and development

References

1. Parker, G. J.; McKiernan, H. E.; Legg, K. M.; Goecker, Z. C., Forensic proteomics. *Forensic Science International: Genetics* **2021,** *54*, 102529.

2. Merkley, E. D.; Wunschel, D. S.; Wahl, K. L.; Jarman, K. H., Applications and challenges of forensic proteomics. *Forensic Sci. Int.* **2019,** *297*, 350-363.

3. Brůžek, J.; Mikšík, I.; Pilmann Kotěrová, A.; Morvan, M.; Drtikolová Kaupová, S.; Santos, F.; Danielisová, A.; Zazvonilová, E.; Maureille, B.; Velemínský, P., Undertaking the biological sex assessment of human remains: The applicability of minimally-invasive methods for proteomic sex estimation from enamel peptides. *Journal of Cultural Heritage* **2024,** *66*, 204-214.

4. Wu, J.; Liu, J.; Ji, A.; Ding, D.; Wang, G.; Liu, Y.; Zhang, L.; Feng, L.; Ye, J., Deep coverage proteome analysis of hair shaft for forensic individual identification. *Forensic Science International: Genetics* **2022,** *60*, 102742.

5. Chu, F.; Mason, K. E.; Anex, D. S.; Jones, A. D.; Hart, B. R., Hair Proteome Variation at Different Body Locations on Genetically Variant Peptide Detection for Protein-Based Human Identification. *Sci. Rep.* **2019,** *9* (1), 7641.

6. Schulte, K. Q.; Hewitt, F. C.; Manley, T. E.; Reed, A. J.; Baniasad, M.; Albright, N. C.; Powals, M. E.; LeSassier, D. S.; Smith, A. R.; Zhang, L.; Allen, L. W.; Ludolph, B. C.; Weber, K. L.; Woerner, A. E.; Freitas, M. A.; Gardner, M. W., Fractionation of DNA and protein from individual latent fingerprints for forensic analysis. *Forensic Science International: Genetics* **2021,** *50*, 102405.

7. Brown, C. O.; Robbins, B. L.; McKiernan, H. E.; Danielson, P. B.; Legg, K. M., Direct seminal fluid identification by protease-free high-resolution mass spectrometry. *J. Forensic Sci.* **2021,** *66* (3), 1017-1023.

8. McKiernan, H. E.; Brown, C. O.; Arantes, L. C.; Danielson, P. B.; Legg, K. M., NextGen Serology: Leveraging Mass Spectrometry for Protein-Based Human Body Fluid Identification. In *Applications in Forensic Proteomics: Protein Identification and Profiling*, American Chemical Society: 2019; Vol. 1339, pp 47-80.

9. Duracova, M.; Klimentova, J.; Fucikova, A.; Dresler, J. Proteomic Methods of Detection and Quantification of Protein Toxins *Toxins (Basel)* [Online], 2018.

10. Kalb, S. R.; Becher, F., Unambiguous Identification of Ricin and Abrin with Advanced Mass Spectrometric Assays. In *Applications in Forensic Proteomics: Protein Identification and Profiling*, American Chemical Society: 2019; Vol. 1339, pp 175-184.

11. Hendy, J., Ancient protein analysis in archaeology. *Science Advances* **2021,** *7* (3), eabb9314.

12. Deshpande, S. V.; Jabbour, R. E.; Snyder, P. A.; Stanford, M. F.; Wick, C. H.; Zulich, A. W., ABOid: A Software for Automated Identification and Phyloproteomics Classification of Tandem Mass Spectrometric Data. *Journal of Chromatography and Separation Techniques* **2011,** *5* (S5), 001-006.

13. Boulund, F.; Karlsson, R.; Gonzales-Siles, L.; Johnning, A.; Karami, N.; AL-Bayati, O.; Ahren, C.; Moore, E. R. B.; Kristiansson, E., TCUP: Typing and characterization of bacteria using bottom-up tandem mass spectrometry proteomics. *Mol. Cell. Proteomics* **2017,** *16* (6), 1052-1063.

14. Alves, G.; Wang, G.; Ogurtsov, A. Y.; Drake, S. K.; Gucek, M.; Suffredini, A. F.; Sacks, D. B.; Yu, Y.-K., Identification of Microorganisms by High Resolution Tandem Mass Spectrometry with Accurate Statistical Significance. *Journal of the American Society for Mass Spectrometry* **2016,** *27*, 194-210.

15. Alves, G.; Wang, G.; Ogurtsov, A. Y.; Drake, S. K.; Gucek, M.; Sacks, D. B.; Yu, Y.-K., Rapid Classification and Identification fo Multiple Microorganisms with Accurate Statistical Significance via High-Resolution Tandem Mass Spectrometry. *Journal of the American Society for Mass Spectrometry* **2018,** *29*, 1721-1737.

16. Alves, G.; Yu, Y.-K., Robust Accurate Identification and Biomass Estimates of Microorganisms via Tandem Mass Spectrometry. *Journal of the American Society for Mass Spectrometry* **2020,** *31* (1), 85-102.

17. Ogurtsov, A.; Alves, G.; Rubio, A.; Joyce, B.; Andersson, B.; Karlsson, R.; Moore, E. R. B.; Yu, Y.-K., MiCId GUI: The Graphical User Interface for MiCId, a Fast Microorganism Classification and Identification Workflow with Accurate Statistics and High Recall. *J. Comput. Biol.* **2024,** *31* (2), 175-178.

18. Alves, G.; Ogurtsov, A. Y.; Porterfield, H.; Maity, T.; Jenkins, L. M.; Sacks, D. B.; Yu, Y.-K., Multiplexing the Identification of Microorganisms via Tandem Mass Tag Labeling Augmented by Interference Removal through a Novel Modification of the Expectation Maximization Algorithm. *Journal of the American Society for Mass Spectrometry* **2024,** *35* (6), 1138-1155.

19. Grenga, L.; Pible, O.; Miotello, G.; Culotta, K.; Ruat, S.; Roncato, M.-A.; Gas, F.; Bellanger, L.; Claret, P.-G.; Dunyach-Remy, C.; Laureillard, D.; Sotto, A.; Lavigne, J.-P.; Armengaud, J., Taxonomical and functional changes in COVID-19 faecal microbiome could be related to SARS-CoV-2 faecal load. *Environ. Microbiol.* **2022,** *24* (9), 4299-4316.

20. Pible, O.; Allain, F.; Jouffret, V.; Culotta, K.; Miotello, G.; Armengaud, J., Estimating relative biomasses of organisms in microbiota using "phylopeptidomics". *Microbiome* **2020,** *8* (1), 30.

21. Jarman, K. H.; Heller, N. C.; Jenson, S. C.; Hutchison, J. R.; Kaiser, B. L. D.; Payne, S. H.; Wunschel, D. S.; Merkley, E. D., Proteomics Goes to Court: A Statistical Foundation for Forensic Toxin/Organism Identification Using Bottom-Up Proteomics. *J. Proteome Res.* **2018,** *17* (9), 3075-3085.

22. Rüther, P. L.; Husic, I. M.; Bangsgaard, P.; Gregersen, K. M.; Pantmann, P.; Carvalho, M.; Godinho, R. M.; Friedl, L.; Cascalheira, J.; Taurozzi, A. J.; Jørkov, M. L. S.; Benedetti, M. M.; Haws, J.; Bicho, N.; Welker, F.; Cappellini, E.; Olsen, J. V., SPIN enables high throughput species identification of archaeological bone by proteomics. *Nature Communications* **2022,** *13* (1), 2458.

23. Yang, H.; Butler, E. R.; Monier, S. A.; Teubl, J.; Fenyö, D.; Ueberheide, B.; Siegel, D., A predictive model for vertebrate bone identification from collagen using proteomic mass spectrometry. *Sci. Rep.* **2021,** *11* (1), 10900.

24. Miura, N.; Okuda, S., Current progress and critical challenges to overcome in the bioinformatics of mass spectrometry-based metaproteomics. *Computational and Structural Biotechnology Journal* **2023,** *21*, 1140-1150.

25. Alves, G.; Ogurtsov, A. Y.; Yu, Y.-K., RAId_DbS: Peptide Identification using Database Searches with Realistic Statistics. *Biol. Direct* **2007,** *2* (1), 25.

26. Padliya, N. D.; Garrett, W. M.; Campbell, K. B.; Tabb, D. L.; Cooper, B., Tandem mass spectrometry for the detection of plant pathogenic fungi and the effects of database composition on protein inferences. *Proteomics* **2007,** *7* (21), 3932-3942.

27. Wunschel, D.; Engelmann, H.; Victry, K.; Clowers, B.; Sorensen, C.; Valentine, N.; Mahoney, C.; Wietsma, T.; Wahl, K., Protein markers for identification of Yersinia pestis and their variation related to culture. *Molecular and Cellular Probes* **2014,** *28* (2–3), 65-72.

28. Mesuere, B.; Debyser, G.; Aerts, M.; Devreese, B.; Vandamme, P.; Dawyndt, P., The Unipept metaproteomics analysis pipeline. *Proteomics* **2015,** *15* (8), 1437-1442.

29. Charlier, P.; Armengaud, J., Did Saint Leonard suffer from Madura foot at the time of death? Infectious disease diagnosis by paleo-proteotyping. *J. Infect.* **2024,** *88* (1), 61-62.

30. Hughes, C.; Ma, B.; Lajoie, G., De Novo Sequencing Methods in Proteomics. In *Proteome Bioinformatics*, Hubbard, S. J.; Jones, A. R., Eds. Humana Press: 2010; Vol. 604, pp 105-121.

31. Muth, T.; Hartkopf, F.; Vaudel, M.; Renard, B. Y., A Potential Golden Age to Come—Current Tools, Recent Use Cases, and Future Avenues for De Novo Sequencing in Proteomics. *Proteomics* **2018,** *18* (18), 1700150.

32. O'Bryon, I.; Jenson, S. C.; Merkley, E. D., Flying blind, or just flying under the radar? The underappreciated power of de novo methods of mass spectrometric peptide identification. *Protein Sci.* **2020,** *29* (9), 1864-1878.

33. Potgieter, M. G.; Nel, A. J. M.; Fortuin, S.; Garnett, S.; Wendoh, J. M.; Tabb, D. L.; Mulder, N. J.; Blackburn, J. M., MetaNovo: An open-source pipeline for probabilistic peptide discovery in complex metaproteomic datasets. *PLOS Computational Biology* **2023,** *19* (6), e1011163.

34. Kleikamp, H. B. C.; Pronk, M.; Tugui, C.; Guedes da Silva, L.; Abbas, B.; Lin, Y. M.; van Loosdrecht, M. C. M.; Pabst, M., Database-independent de novo metaproteomics of complex microbial communities. *Cell Systems* **2021,** *12* (5), 375-383.e5.

35. Lee, J.-Y.; Mitchell, H. D.; Burnet, M. C.; Wu, R.; Jenson, S. C.; Merkley, E. D.; Nakayasu, E. S.; Nicora, C. D.; Jansson, J. K.; Burnum-Johnson, K. E.; Payne, S. H., Uncovering Hidden Members and Functions of the Soil Microbiome Using De Novo Metaproteomics. *J. Proteome Res.* **2022,** *21* (8), 2023-2035.

36. Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G., PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2003,** *17* (20), 2337-2342.

37. Ma, B., Novor: Real-Time Peptide de Novo Sequencing Software. *Journal of the American Society for Mass Spectrometry* **2015,** *26* (11), 1885-1894.

38. Yilmaz, M.; Fondrie, W.; Bittremieux, W.; Oh, S.; Noble, W. S., De novo mass spectrometry peptide sequencing with a transformer model. In *Proceedings of the 39th International Conference on Machine Learning*, Kamalika, C.; Stefanie, J.; Le, S.; Csaba, S.; Gang, N.; Sivan, S., Eds. PMLR: Proceedings of Machine Learning Research, 2022; Vol. 162, pp 25514--25522.

39. Muth, T.; Renard, B. Y., Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Briefings in Bioinformatics* **2018,** *19* (5), 954-970.

40. Dasari, S.; Chambers, M. C.; Slebos, R. J.; Zimmerman, L. J.; Ham, A.-J. L.; Tabb, D. L., TagRecon: High-Throughput Mutation Identification through Sequence Tagging. *J. Proteome Res.* **2010,** *9* (4), 1716-1726.

41. Devabhaktuni, A.; Lin, S.; Zhang, L.; Swaminathan, K.; Gonzalez, C. G.; Olsson, N.; Pearlman, S. M.; Rawson, K.; Elias, J. E., TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. *Nat. Biotechnol.* **2019,** *37* (4), 469-479.

42. Sahl, J. W.; Vazquez, A. J.; Hall, C. M.; Busch, J. D.; Tuanyok, A.; Mayo, M.; Schupp, J. M.; Lummis, M.; Pearson, T.; Shippy, K.; Colman, R. E.; Allender, C. J.; Theobald, V.; Sarovich, D. S.; Price, E. P.; Hutcheson, A.; Korlach, J.; LiPuma, J. J.; Ladner, J.; Lovett, S.; Koroleva, G.; Palacios, G.; Limmathurotsakul, D.; Wuthiekanun, V.; Wongsuwan, G.; Currie, B. J.; Keim, P.; Wagner, D. M., The Effects of Signal Erosion and Core Genome Reduction on the Identification of Diagnostic Markers. *mBio* **2016,** *7* (5).

43. Merkley, E. D.; Jenson, S. C.; Arce, J. S.; Melville, A. M.; Leiser, O. P.; Wunschel, D. S.; Wahl, K. L., Ricin-like proteins from the castor plant do not influence liquid chromatography-mass spectrometry detection of ricin in forensically relevant samples. *Toxicon* **2017,** *140* (Supplement C), 18-31.

44. Pfrunder, S.; Grossmann, J.; Hunziker, P.; Brunisholz, R.; Gekenidis, M.-T.; Drissner, D., Bacillus cereus Group-Type Strain-Specific Diagnostic Peptides. *J. Proteome Res.* **2016,** *15* (9), 3098-3107.

45. Vizcaíno, J. A.; Csordas, A.; del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q.-W.; Wang, R.; Hermjakob, H., 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **2016,** *44* (D1), D447-D456.

46. Deutsch, E. W.; Csordas, A.; Sun, Z.; Jarnuczak, A.; Perez-Riverol, Y.; Ternent, T.; Campbell, D. S.; Bernal-Llinares, M.; Okuda, S.; Kawano, S.; Moritz, R. L.; Carver, J. J.; Wang, M.; Ishihama, Y.; Bandeira, N.; Hermjakob, H.; Vizcaíno, J. A., The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* **2017,** *45* (D1), D1100-D1106.

47. Payne, S. H.; Monroe, M. E.; Overall, C. C.; Kiebel, G. R.; Degan, M.; Gibbons, B. C.; Fujimoto, G. M.; Purvine, S. O.; Adkins, J. N.; Lipton, M. S.; Smith, R. D., The Pacific Northwest National Laboratory library of bacterial and archaeal proteomic biodiversity. *Scientific Data* **2015,** *2*, 150041.

48. Ehling-Schulz, M.; Lereclus, D.; Koehler Theresa, M., The Bacillus cereus Group: Bacillus Species with Pathogenic Potential. *Microbiol Spectr* **2019,** *7* (3), 10.1128/microbiolspec.gpp3-0032-2018.

49. Liu, Y.; Lai, Q.; Göker, M.; Meier-Kolthoff, J. P.; Wang, M.; Sun, Y.; Wang, L.; Shao, Z., Genomic insights into the taxonomic status of the Bacillus cereus group. *Sci. Rep.* **2015,** *5* (1), 14082.

50. Schoch, C. L.; Ciufo, S.; Domrachev, M.; Hotton, C. L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; McVeigh, R.; O'Neill, K.; Robbertse, B.; Sharma, S.; Soussov, V.; Sullivan, J. P.; Sun, L.; Turner, S.; Karsch-Mizrachi, I., NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database : the journal of biological databases and curation* **2020,** *2020*, baaa062.