Journal of proteome research

pubs.acs.org/jpr

Article

MARLOWE: An Untargeted Proteomics, Statistical Approach to Taxonomic Classification for Forensics

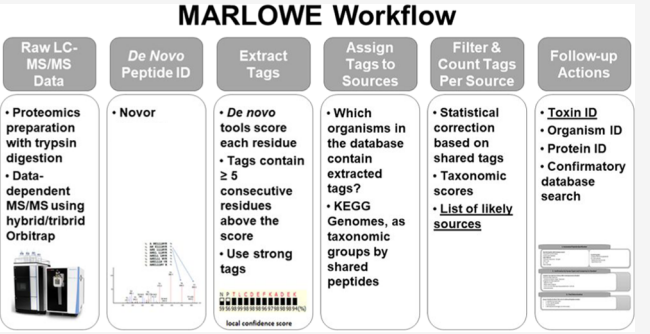
Fanny Chu,*,⊥ Sarah C. Jenson,⊥ Anthony S. Barente, Natalie C. Heller, Eric D. Merkley, and Kristin H. Jarman

Cite This: https://doi.org/10.1021/acs.jproteome.3c00477

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: General proteomics research for fundamental science typically addresses laboratory- or patient-derived samples of known origin and composition. However, in a few research areas, such as environmental proteomics, clinical identification of infectious organisms, archeology, art/cultural history, and forensics, attributing the origin of a protein-containing sample to the organisms that produced it is a central focus. A small number of groups have approached this problem and developed software tools for taxonomic characterization and/or identification using bottom-up proteomics. Most such tools identify peptides via database search, and many rely on organism-specific peptides as markers. Our group recently introduced MARLOWE, a software tool for taxonomic characterization of unknown samples based on *de novo* peptide identification and signal-erosion-resistant strong peptides, which are shared peptides distributed in a taxonomy-dependent manner. In the current work, we further characterize the utility of MARLOWE using publicly available proteomics data from forensically-relevant samples. MARLOWE characterizes samples based on their protein profile, and returns ranked organism lists of potential contributors and taxonomic scores based on shared strong peptides between organisms. Overall, the correct characterization rate ranges between 44 and 100%, depending on the sample type and data acquisition parameters (with lower numbers associated with lower-quality data sets). MARLOWE demonstrates successful characterization of true contributors and close relatives, and provides sufficient specificity to distinguish certain microbial species. MARLOWE demonstrates its ability to provide insight into potential taxonomic sources for a wide range of sample types without prior assumptions about sample contents. This approach can find utility in forensic science and also broadly in bioanalytical applications that utilize proteomics approaches for taxonomic characterization.



KEYWORDS: *forensic proteomics, bioforensics, liquid chromatography-tandem mass spectrometry, strong peptides, de novo sequence tags*

INTRODUCTION

Source attribution of biological samples remains an important part of bioforensics, yet there are few tools that utilize the protein profile to make this determination when the sample is completely unknown. Classification and identification for forensic applications necessitate use of proteomics methods when DNA is absent, compromised, or when proteins—not nucleic acids—are the agent of interest, such as for protein toxins. For example, characterization of the protein toxin ricin from castor seeds is a common application in forensic proteomics, and extensive work has been performed to establish criteria for identifying ricin.[1−4] Other applications of class-based forensic identification include identification of other toxins[5,6] and species identification, such as for the presence of pathogens.[7−10] Much of the work towards classification in bioforensics using proteomics approaches has utilized conventional proteomics data acquisition and analysis approaches. However, these conventional approaches make assumptions about the unknown samples, which may not always be justified nor forensically defensible.

Typical bottom-up proteomics approaches, such as for fundamental science studies, rely on database search to identify peptides, where the database contains protein sequences from organisms known to be present in the sample, e.g., because the samples were generated as part of the experiment. This method of database selection assumes that the user-defined database contains protein sequences from the true contributor(s) of the sample, which may not always be the case,[11] especially in forensic samples. Other applications, such as metaproteomics, utilize a broader database that typically contains the protein

Received: August 1, 2023
Revised: November 7, 2024
Accepted: January 21, 2025
Published: February 3, 2025

sequences of hypothesized contributors where possible. After performing the database search with a user-defined database, identification of "unique" peptides attributed to specific organisms, where "unique" peptides are only observed in a specific, single organism, then allows for taxonomic source attribution. However, this method of determining taxonomic origin suffers from signal erosion, in which continued sequencing of organism genomes reduces the probability that "unique" peptides will remain as such over time. Pfrunder and co-workers illustrate an example of signal erosion in their approach to identify characteristic peptides for distinction of *Bacillus cereus* group members.[12] Between May and October 2015, characteristic peptides for each of the 7 species examined were reduced by roughly 46% on average, and for some species, the actual number of "unique" peptides was <8.[12] In this case, use of "unique" peptides to identify members within this superspecies may not be statistically robust. For forensic applications where samples may be complex mixtures, more robust methods are needed to confidently characterize samples of unknown taxonomic origin.

Although organism classification and identification for forensically relevant use cases has a distinct goal from metaproteomics, some published metaproteomics approaches address the limitations associated with identifying unique markers from organism subsets. For example, MetaProteomeAnalyzer (MPA)[14] and UniPept Metaproteomics Analysis provide taxonomic source information from peptides identified in a sample. UniPept Metaproteomics Analysis matches sample peptides to *in silico* tryptic peptides in their database, which then map to proteomes of organisms in the UniProt database, to make lowest common ancestor determinations.[13] However, assignment to lowest common ancestor may provide limited taxonomic specificity, and their lowest common ancestor approach relies on equal weighting of all peptides, which inflates the weight of nonspecific peptides. MPA provides various approaches to infer taxonomy from peptides by applying different combinations of peptide and protein similarity rules and accounting for peptide sequence variations, and returns either the lowest common ancestor or most specific taxonomy from the set of identified peptides,[14,15] enabling greater taxonomic specificity compared to UniPept's algorithm. However, despite moving away from targeting "unique" markers, both MPA and UniPept rely on broad database searches.

Other tools have more specifically focused on taxonomic characterization. TCUP, a workflow for characterizing bacterial mixtures presented by Boulund et al., utilizes a similar strategy but then only considers peptides with lowest common ancestor determinations at the strain level for taxonomic source assignment in samples.[16] As described above, using only taxonomically distinct peptide markers suffers from signal erosion. A leading tool in organism identification, MiCId, uses shared peptides, with their contribution to organism identification weighted by the inverse of the number of clusters in which that peptide appears.[11,17,18] A recent software tool called SPIN, designed for use with mammalian bone fragments, uses a database of only bone proteins. Identified peptides are scored against a series of pairwise species sequence differences, and the species that wins the most pairwise comparisons is the overall winner.[19] Finally, the phylopeptidomics approach uses "taxon-spectrum matches," assigning peptides to all the taxa to which they match.[20] However, despite moving away from targeting "unique" peptide markers to determine taxonomic origin, these tools still require peptide identification through database searching, which as mentioned above, is a type of targeted data analysis that assumes that the source organism proteomes are included in the database, which may not always be true for complex mixtures such as those potentially encountered in forensic applications. As such, there is a need for an untargeted organism identification/classification tool to characterize unknown samples that can be encountered in forensic evidence.

We previously demonstrated utility of an organism identification/classification tool to characterize unknown samples in an untargeted manner, which we call MARLOWE (after the fictional detective introduced by author Raymond Chandler) (Jenson et al.).[21] MARLOWE is intended to provide insight into the taxonomy/identity of organisms that contribute protein to unknown forensic samples, including but not limited to whole organisms. The goal of MARLOWE is to provide sufficient taxonomic resolution to provide investigative leads and identify appropriate follow-on studies (for example, confirmatory database searches, targeted mass spectrometry assays, or toxin detection). Its purpose is thus similar, but not identical, to other proteotyping tools such as those described above. MARLOWE matches *de novo* peptide sequence tags, which include peptide precursor information, to *in silico* tryptic peptides from the KEGG database, assigns these matches to organisms through protein inference, and weights the tag-peptide matches based on how frequently these peptides are found in multiple organisms. To our knowledge, *de novo* peptide sequence information has not been used towards characterizing microbial samples, but can be a valuable, unbiased source of proteomic information for taxonomic origin attribution. Johnson and co-workers[22] have used *de novo* peptide identification to evaluate the suitability of a database from a closely related organism, but did not address the problem of assigning taxonomic origin of a completely unknown sample. As a proof-of-concept study, MARLOWE successfully characterized pure and binary mixtures of bacterial cultures (Jenson et al.).[21] Here, we extend characterization to a wide range of forensically-relevant sample types.

This research aims to benchmark MARLOWE's characterization performance for analysis of forensically-relevant biological samples and provide insight into interpreting such taxonomic characterizations for different sample types. Specifically, the ability for MARLOWE to correctly characterize multiple-contributor mixtures, characterize components in ground seed extracts, demonstrate characterization specificity for closely related mammalian samples, and characterize contributors in degraded samples such as archeological and historical artifacts were examined using publicly available mass spectrometry data sets and an in-house data set. We find that MARLOWE correctly characterizes most contributors of microbial samples at the species level and, though with less accuracy, can characterize closely related mammalian samples. Successful characterization hinges on having high-quality, intact samples and high mass spectrometry data quality, which may be slightly more challenging for degraded and ancient samples. MARLOWE further provides insight into proteome diversity of organisms and their close relatives, which may not be reflected in their assigned taxonomy.

https://doi.org/10.1021/acs.jproteome.3c00477
J. Proteome Res. XXXX, XXX, XXX−XXX

METHODS

Ground Seed Extract Preparation and Mass Spectrometry Data Acquisition

Castor seeds (*Ricinus communis*), jequirity peas (*Abrus precatorius*), soybeans (*Glycine max*), and peanuts (*Arachis hypogaea*) were ground into mash using a glass beaker and a microspatula as described in Wunschel et al.[23] Mash was extracted by grinding with a micropestle in the presence of phosphate buffered saline (PBS) buffer. Extracts were centrifuged at 4 °C to remove insoluble material and oil, and heated at 100 °C for 1 h to inactivate toxins. Samples were prepared and digested using a proteomics protocol described in Merkley et al.[1] Briefly, following inactivation, extracts were denatured and reduced using 8 M urea and 5 mM dithiothreitol at 60 °C for 1 h, and alkylated with 15 mM iodoacetamide for 1 h in the dark at 37 °C. Samples were digested overnight at 37 °C with trypsin at a 1:50 w/w trypsin/protein ratio in the presence of 1 mM calcium chloride and desalted using solid phase extraction with C18 resin.

Protein digests were separated on a Waters nanoAcquity liquid chromatograph (Milford, MA) using in-house fused silica capillary columns packed with Jupiter C18 stationary phase (Phenomenex, Torrance, CA). Five μL injections were performed to load samples onto the trapping column (4 cm x 150 μm i.d., 5 μm particle size) at a flow rate of 3 μL/min. Chromatographic separation was achieved at a flow rate of 300 nL/min on the analytical column (70 cm x 75 μm i.d., 360 μm o.d., 3 μm particle size) using mobile phases A (0.1% formic acid in water) and B (0.1% formic acid in acetonitrile), and the following gradient: 2% B at 0 min, 8% B at 2 min, 12% B at 20 min, 30% B at 75 min, 45% B at 97 min, 95% B at 100 min, 95% B at 110 min, 1% B at 115 min, and 1% B at 150 min, followed by a "sawtooth" gradient wash to mitigate carryover.

Mass spectra were acquired on a Q Exactive HF (Thermo Scientific, San Jose, CA) mass spectrometer in data-dependent tandem mass spectra (MS/MS) mode. Nanoelectrospray ionization was performed at 2300 V and mass spectrometry data were collected between 15 and 115 min of the chromatographic gradient. Full MS spectra were acquired at a resolution of 60,000 and the top 12 most abundant ions were selected for MS/MS. Precursor ions were fragmented at a normalized collision energy of 30 using higher-energy collision induced dissociation (HCD). Tandem mass spectra were acquired at a resolution of 15,000 and an isolation window of 2 Da. Dynamic exclusion was set to 30 s. Raw mass spectra were then converted to MGF format using ThermoRawFileParser.

Publicly Available Raw Mass Spectrometry Data

Raw spectrometry data files were downloaded from ProteomeXchange, from the following projects: PXD004321, PXD018933, PXD008103, and PXD001029, and can be accessed via the ProteomeXchange online repository[24,25] (http://proteomecentral.proteomexchange.org). Metadata for data files in PXD004321 were obtained from Boulund and coworkers (personal communication). These data sets were converted to MGF format using ThermoRawFileParser.

De Novo Peptide Sequencing

MARLOWE relies on output from *de novo* peptide sequencing as input. From tandem mass spectrometry data, *de novo* peptide sequencing was performed using Novor.[26] Each data set contains different types of samples and matrices, and were acquired on different Orbitrap instruments. Appropriate *de novo* sequencing parameters, such as fragment mass error tolerance and fixed and variable modifications, were set to align with the different sample type, matrix, and data acquisitions per data set (Supporting Table S1). Output from *de novo* peptide sequencing via Novor was obtained for each mass spectrometry data file as input to MARLOWE. Filtering of Novor peptide-spectrum matches to obtain sequence tags is described in detail in the next section, but in brief, a tag is defined as a unit with at least 5 consecutive residues with local confidence scores of 80 or greater, from a peptide-spectrum match with an average local confidence score of 50 or greater.

MARLOWE Architecture and Analysis

The statistical concepts underlying MARLOWE have been described in detail in Jenson et al.[21] Briefly, MARLOWE includes an SQL database containing in-silico-generated tryptic peptides from the proteomes of all organisms in the KEGG Genome database Release 91.0 (https://www.genome.jp/kegg/genome/; downloaded July 1, 2019) and takes in output from *de novo* peptide sequencing. These outputs are formatted into lists of *de novo* sequence tags to use as query input to the SQL database. Results of these queries manifest as ranked lists of organisms, with associated taxonomic scores, that are potential contributors based on matching *de novo* tags to strong peptides belonging to each organism.

From output files generated by Novor, *de novo* peptides were filtered to include only highly-confident regions known as sequence tags, typically those peptides with an overall (average) local confidence score of 50 and at least 5 consecutive residues with amino acid local confidence scores greater than 80. Note that local confidence scores range between 0 and 100, with higher scores denoting high confidence in residue assignment.[26] Supporting Table S1 includes specific threshold values for each data set for MARLOWE.

Two concepts underlying MARLOWE are strong peptides and taxonomic groups. Strong peptides are *in silico* fully tryptic peptides without missed cleavages that are found in fewer than approximately 5% of organisms in the database, as determined at the genus level.[8] Strong peptides occur much more frequently in one genus compared to all other genera. These strong peptides could be unique peptides, which are only found in a single organism, or they could be shared among organisms, likely from the same genus, but occur in less than approximately 5.3% of organisms' proteomes. (Note that this value is a result of applying a 0.95 threshold value to the peptide strength equation.) Jarman et al.[8] and Jenson et al.[21] provide much more detailed explanations on the theory of the strong peptide calculation. Here, the pairwise peptide strength adaptation is utilized, as described in Jenson et al.[21] This implementation considers the frequency of a peptide within each genus, and defines a peptide as strong if (1) its presence within the genus where it occurs most frequently is at least 95% and (2) it is found in 5% or fewer species at the genus where the peptide occurs second most often. At a high level, peptide strength is a probabilistic concept that while a newly sequenced organism's proteome might contain an otherwise unique peptide, thus rendering the peptide not unique, it is much less likely that the presence of a strong peptide in that same newly sequenced organism's proteome would result in that strong peptide being present at 5% or greater of all sequenced organisms' proteomes; the strong peptide remains strong. When present and detected, strong peptides serve as

https://doi.org/10.1021/acs.jproteome.3c00477
J. Proteome Res. XXXX, XXX, XXX−XXX

Table 1. Characterization Summary of MARLOWE for Each Dataset with Known True Contributors

data set	number of samples	characterization rate (% top 1/top2/top5)[a] family	genus	species	taxonomic group	data set reference
TCUP	42	pure: 100/−/− mixtures: 72/78 (top 4/top 5)	Pure: 100/−/− Mixtures: 56/78 (top 4/top 5)	Pure: 100/−/− Mixtures: 50/78 (top 4/top 5)	Pure: 100/−/− Mixtures: 72/78 (top 4/top 5)	Boulund et al.[16]
primate teeth	20	44/75/81	44/56/56	44/56/56	56/81/81	Froment et al.[27]
seeds	11	100/−/−	100/−/−	100/−/−	100/−/−	

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

evidence for presence of the organism or organisms that contain them, regardless of the formal taxonomy of those organisms. Aggregated strong peptide counts become strong evidence for their source organisms. Even though strong peptide presence criteria were defined at the genus level, this categorization of strong peptides does not align with any formal taxonomic ontology (in contrast to family-specific peptides or genus-specific peptides), as the set of strong peptides is determined based on the set of all theoretical peptides from considered organisms' proteomes. MARLOWE's taxonomic groups comprise organisms that all share at least 40% of their theoretical peptides with each other, formed using a leader clustering algorithm based on number of shared tryptic peptides. MARLOWE's taxonomic groups do not necessarily align with any specific taxonomic ontology (e.g., family, species), as they are based on proteomic and thus genomic/phylogenetic similarity. This grouping presents an alternative to formal taxonomic ontology, of which its issues in relation to genomic similarity are well-known.

The list of *de novo* sequence tags, which consists of pairs of tag sequence and peptide precursor monoisotopic mass, is used as a query and matched to all *in silico* peptides in the SQL database within a mass error tolerance of ±15 ppm. To limit protein inference, candidate proteins inferred from these tag-peptide matches are filtered to include only those that have had a minimum of two peptide assignments, one of which must be from a strong peptide. The tag-peptide matches are then pruned to include only those whose peptides are strong, as defined at the genus level, and then assigned to organisms whose proteomes contain those *in silico* strong peptides. An optional tag-strong peptide match filter can be applied at this stage to remove matches to peptides in contaminants, given as a fasta file. After these filters, tag-strong peptide matches are tabulated for each organism and taxonomic group, respectively, and this latter count is transformed into taxonomic score.

Taxonomic score, the primary metric generated by MARLOWE, is a weighted count of the number of tag-strong peptide matches assigned to each taxonomic group. The weighting, using a non-negative least-squares model, exploits a reward/penalty function based on the similarity between taxonomic groups using shared strong peptides, such that similar taxonomic groups should have similar number of tag-strong peptide matches (Jenson et al.)[21] Scores that align with this correlation are positively weighted (elevated), and negatively weighted (diminished) if not. For each sample, MARLOWE returns a list of organisms with associated number of tag-strong peptide matches and taxonomic score.

Data Analysis

From organism lists returned by MARLOWE, results were further filtered to include only taxa that received a minimum of 2 tag-strong peptide matches. For ease of comparison across samples, taxonomic score was then normalized to the total taxonomic score per sample, such that normalized score ranges between 0 and 1. Organisms at the species level were then ranked by decreasing taxonomic score and decreasing tag-strong peptide matches. Ranked organism lists were then compared across samples to assess MARLOWE's characterization performance.

RESULTS AND DISCUSSION

In this paper, we seek to evaluate an organism identification/classification approach that utilized *de novo* peptide identification combined with the concept of peptide strength. *De novo* peptide identification is well-understood, and has been used somewhat in the related field of metaproteomics.[13,14] However, peptide strength is newer, and for readers not familiar with the relevant publications, we briefly describe it here. Peptide strength describes the nonuniformity of occurrence of a peptide sequence across (in this case) genera within a database of organisms using weight-of-evidence statistical concepts that are common in forensics.[8,21] Strong peptides are not simply taxon-specific peptides; they are shared peptides. However, because they are shared in a nonuniform way across taxa, they serve as pointers, not to a single organism or taxon, but to a group of taxa. The MARLOWE algorithm aggregates this data by evaluating all the organisms to which all observed strong peptides point and providing a ranking.

We demonstrate MARLOWE's performance using publicly available data sets that represent some of the scope of samples that could be encountered in forensic evidence. Samples of interest include multiple-contributor mixtures, mammalian samples, and archeological and historical artifacts that have met with some degree of degradation. To that end, we examined five data sets to address the following questions:

1. Can MARLOWE characterize components from multiple-contributor mixtures?
2. Can MARLOWE characterize components in archeological and historical artifacts?
3. How does MARLOWE's performance compare between microbial and mammalian samples?
4. Can MARLOWE characterize components extracted from plant material, specifically, ground seeds?

Using *de novo* peptide sequencing as query inputs, MARLOWE returns a list of potential organisms for each sample and associated taxonomic scores. Organisms are then ranked by highest taxonomic score and the greatest number of tag-strong peptide matches. Potential organisms and their ranks can then be compared to true contributors for quantifying characterization success. Table 1 summarizes MARLOWE's characterization performance for each data set,

https://doi.org/10.1021/acs.jproteome.3c00477
J. Proteome Res. XXXX, XXX, XXX−XXX

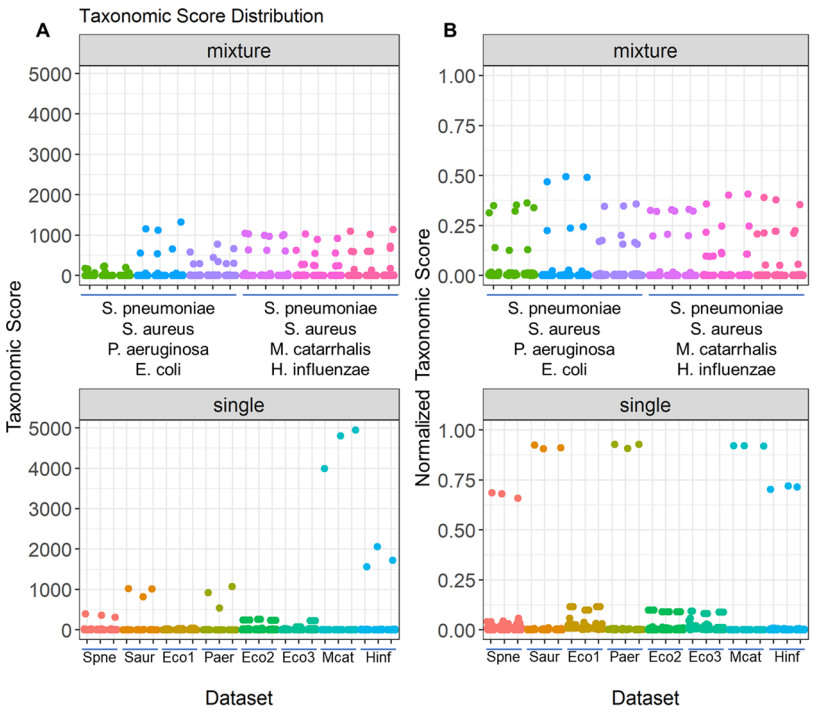


Figure 1. (A) Raw and (B) normalized taxonomic score distributions from ranked lists returned by MARLOWE, for pure (single) bacterial samples and 4-species mixtures. Different sample compositions (i.e., microbial species in single samples and ratios of different microbial species in mixtures) are delineated by color. Taxonomic score distributions within each sample enable distinction of pure samples and mixtures, as the taxonomic score associated with the true contributor for pure samples is substantially higher than scores for all other taxa. In contrast, taxonomic scores for true contributors in most mixtures are lower, yet still distinct from other taxa.

at the family, genus, species, and taxonomic group levels, where true contributors are known. Family-level characterization was included since the KEGG database, on which MARLOWE's SQL database relies, may not contain all sequenced organisms to date, including some at the genus level. Characterization rate is defined as correct characterization of the true contributor as the top-ranked organism, and rates are listed for correct characterization at the top 1, 2, and 5 ranks. For mixtures, characterization rate is the correct characterization of all true contributors as the top 4- and 5-ranked organisms. Deeper examination of MARLOWE's characterization performance for each data set, including those where true contributors are not known, is described in the sections below, through comparisons of raw taxonomic scores, normalized taxonomic scores, and tag-strong peptide matches.

Analysis of Multiple-Contributor Samples: TCUP

Previous work describing MARLOWE's characterization performance focused on single-contributor and simulated binary mixtures of bacterial cultures (Jenson et al.).[21] Here, we challenged MARLOWE to correctly characterize all components of 4-species mixtures. Complex mixtures from multiple taxonomic sources may be encountered in forensic evidence, such as in urine, and the goal of the forensic analysis may be to identify specific contributors in complex mixtures. This data set comprises three different ratios of two sets of 4-species mixtures. The first set of mixtures include *Escherichia coli*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Pseudomonas aeruginosa*, while the second set, which simulates the environment in a coinfected respiratory tract sample, contains *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Moraxella catarrhalis*.[16] We compare MARLOWE's performance in correct characterization of pure and complex mixtures, and discuss interpretation of taxonomic scores, including the effects of different sample compositions and concentrations on taxonomic scores and distinction of potential contributors from all other taxa.

MARLOWE correctly characterized all components in pure samples as the top-ranked organism and within the top 5 organisms in most 4-species mixtures at the species level. As expected, for pure bacterial samples, the correct organism was returned as the top hit, with the highest taxonomic score and the greatest number of tag-strong peptide matches. MARLOWE returned all 4 contributors in most 1:1:1:1, 1:2:2:4, and 4:2:2:1 mixtures as within the top 5 hits (Table 1). However, even when true contributors are correctly characterized as the top-ranked organisms, there are some differences in their taxonomic scores, which depend on sample composition and concentration.

Pure (single-species) samples typically yield higher taxonomic scores for the top-ranked organism compared to mixtures owing to higher concentrations, but scores also depend on the potential source organisms. Figure 1 displays raw and normalized taxonomic score distributions for each sample. Whereas most top-ranked sources in pure samples have taxonomic scores ranging between 500 and 5000 (normalized taxonomic score >0.60), scores for top-ranked organisms in mixtures are lower, ranging between 500 and 1500 (normalized score between 0.10 and 0.50) (Figure 1). However, among the different true contributors, taxonomic scores for *E. coli* in both pure samples and mixtures are much lower. Though still included within the top 6 for the majority of *E. coli*-containing mixture replicates, taxonomic scores and the number of tag-strong peptide matches to *E. coli* are generally low (i.e., on average, 0.008 ± 0.005 (s.d.) normalized

https://doi.org/10.1021/acs.jproteome.3c00477
J. Proteome Res. XXXX, XXX, XXX−XXX

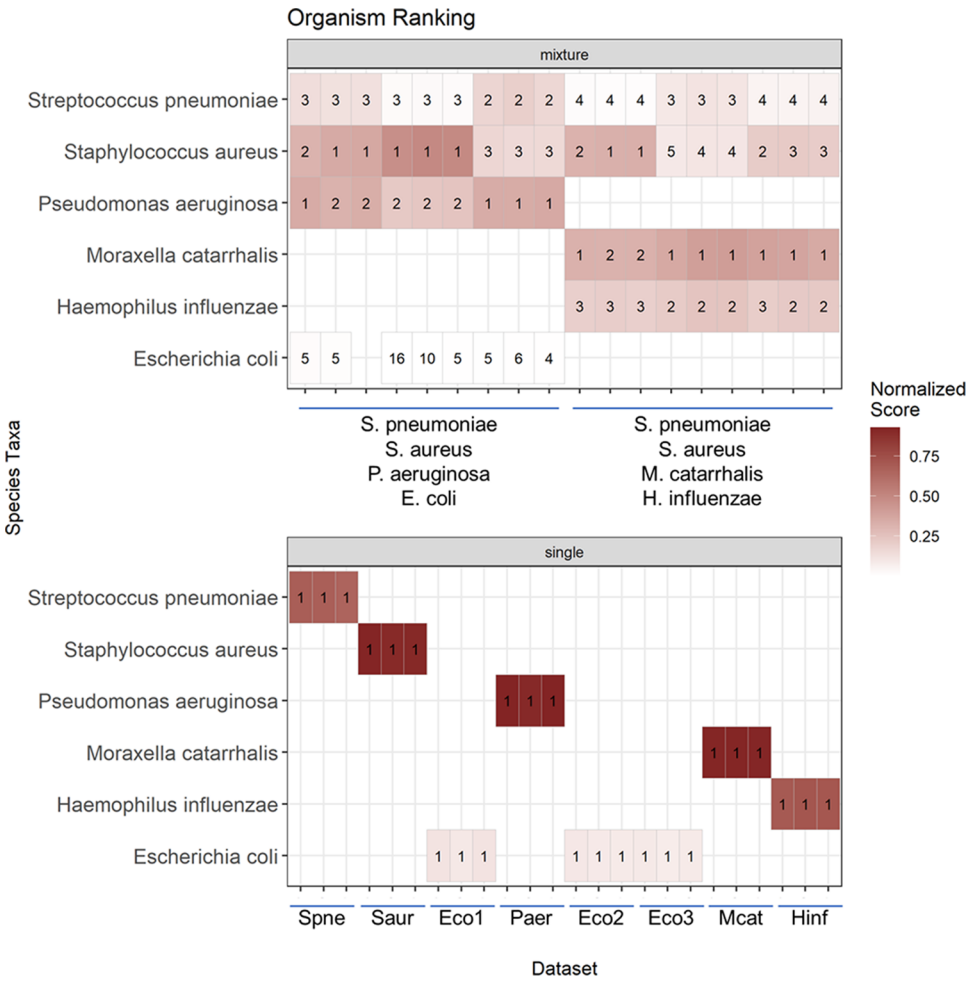


Figure 2. Heatmap of taxonomic scores (color scale) and ranking (numbers) for correctly characterized organisms returned by MARLOWE for pure (single) bacterial samples and mixtures. MARLOWE correctly characterized organisms in pure samples as the top hit, and within the top 5 organisms for most 4-species mixtures.

taxonomic score and 4 ± 2 (s.d.) matches) (Figure 2). This is apparent when comparing the normalized taxonomic scores for *E. coli* across pure samples and in mixtures. On average, taxonomic scores for *E. coli* are 190.8 ± 90.6 (s.d.) for pure samples and 11.6 ± 3.6 (s.d.) for mixtures. However, for other taxa such as *P. aeruginosa*, taxonomic scores in both pure samples and mixtures are substantially higher (i.e., on average, 874.6 ± 216.6 (s.d.) and 557.8 ± 183.8 (s.d.) for pure samples and mixtures, respectively). Lower taxonomic scores for *E. coli* compared to other taxa result from having few strong peptides that are representative of the taxonomic group containing *E. coli*, owing to the presence of a large number of highly similar proteomes from related genera in other taxonomic groups. Because taxonomic scores derive from tag-strong peptide matches, having few strong peptides for *E. coli* limits the number of possible tag-strong peptide matches. However, despite the lower taxonomic score, MARLOWE returns *E. coli* as a potential contributor in all pure samples and all but one mixture, with a minimum taxonomic score of 7.9.

Notably, taxonomic scores can provide insight into whether the sample contains one component or derives from multiple contributors, and comparison of score distributions within each sample permit distinction of contributors and non-contributors. Thus, far, we have not applied any thresholds to taxonomic score to distinguish hits to potential contributors from all other taxa. MARLOWE returns a ranked organism list based on all *de novo* sequence tags that match to strong peptides per organism, and organisms have been assembled into taxonomic groups. As such, this list comprises not only true contributors, but also any taxa that share nonspecific tag-strong peptide matches. But because taxonomic score depends on the number of tag matches to strong peptides and the relatedness of taxa, organisms that are not true contributors will have low numbers of tag-strong peptide matches and low taxonomic scores, compared to those for true contributors. In this manner, the distribution for normalized taxonomic scores in pure samples is such that there is a distinctly high taxonomic score for the true contributor and low taxonomic scores for all other taxa (Figure 1). In contrast, higher taxonomic scores for each contributor in the mixtures are observed, and when normalized, the taxonomic scores of each of the contributors is smaller than their pure sample counterparts, yet still distinct from all other noncontributor taxa (Figure 1).

Given that MARLOWE returns low taxonomic scores for *E. coli* as a contributor, as discussed above, the distinction between true contributor and noncontributors for *E. coli* is slightly more ambiguous, even in pure *E. coli* samples (Figure 1). Further, some noncontributors may be highly similar to *E. coli* and organized into the same taxonomic group as *E. coli*, resulting in identical taxonomic scores. Indeed, this is the case

https://doi.org/10.1021/acs.jproteome.3c00477
J. Proteome Res. XXXX, XXX, XXX−XXX

for pure *E. coli* samples, where the second ranked organism in all replicates was *Shigella dysenteriae* (specifically strain 1617), an organism in the same Enterobacteriaceae family and the same MARLOWE taxonomic group as *E. coli*. The proteomic similarity between *Shigella* species and *E. coli* has also been noted by Boulund and co-workers, who removed sequences of *Shigella* species from their organism reference database within TCUP.[16] This selection of organisms in the reference database enables better distinction of *E. coli* in samples, but introduces bias into the data analysis process. In contrast, MARLOWE retains these sequences and hits to both *S. dysenteriae* and *E. coli* can be observed, albeit with lower taxonomic scores than for other organisms. In these cases, characterization of organisms at the taxonomic group level provides valuable insight into taxonomic sources of unknown samples that can be used in subsequent data analysis steps, such as selecting relevant protein databases for database searching of tandem mass spectral data, toward more confirmatory source attribution.

Compared to TCUP, very limited quantitative information can be obtained for components in these mixtures from MARLOWE because organism scores are identical for organisms in the same taxonomic group and are weighted by peptide strength among organisms, which dissociate any abundance information specific to each organism. However, with 100% correct characterization, MARLOWE outperforms the true positive rates reported by Boulund and co-workers on replicates of single-species *E. coli*, *P. aeruginosa*, *S. aureus*, and *S. pneumoniae* cultures.[16] Further, correct characterization for mixtures other than those containing *E. coli* was also very successful, demonstrating MARLOWE's ability as a qualitative tool to characterize complex mixtures.

Analysis of Archeological and Historical Artifacts: Danish Peat Bog Burial Site Artifacts, Plague Manuscript Pages

We examined MARLOWE's performance on the Danish peat bog and plague papers data sets, which derive from samples that have undergone some degree of degradation. These data sets represent yet another sample type that could be encountered in forensics, that is, unknown samples with some degradation that are of historical and/or archeological value and whose true contributors are not known. The Danish peat bog data set contains archeological skin samples obtained from garments discovered at three different peat bog burial sites.[28] Characterization of the organisms (sources) that were used to create these garments provides insight into the agricultural significance and evolution of livestock and domesticated species over time. The plague papers data set, on the other hand, comprises protein extracts lifted from death registries in Milan created during the bubonic plague in 1630, which was caused by the etiological agent *Yersinia pestis*.[29]

Analysis of these data sets was intended to provide insight into taxonomic origin of truly unknown samples, however, MARLOWE returned lists of potential contributors with extremely low taxonomic scores (on average, 3.4 ± 3.1 (s.d.)). Of notable results for the plague papers data set, MARLOWE returned hits to *Mus musculus* in two samples (Supporting Figure S1), which align with results described by D'Amato and co-workers.[29] However, *Y. pestis* was not returned as a potential organism in any of the samples (Supporting Figure S1), a departure from reported results. In this case, MARLOWE's characterization may have been adversely affected by low numbers of *de novo* peptides and of highly confident *de novo* tags. Deeper examination of raw mass spectrometry data indicated that spectral quality was low, as quantified by local confidence scores of *de novo* peptides generated by Novor (Supporting Figure S2). Local confidence scores measure the confidence of *de novo* peptide sequencing, and highly confident sequence tags are needed for MARLOWE. For the plague papers data set, it is likely that low spectral quality stems from the insufficient time spent acquiring each MS/MS scan during data acquisition, as the top 40 most intense peptides were selected for fragmentation per precursor ion scan,[29] as opposed to the more typical top 12−20 scans for the Thermo Fisher Orbitrap Fusion instrument used in this study. We note that the top 40 acquisition method was performed using an ion trap mass analyzer in the Orbitrap Fusion instrument, and the mass resolution that is achievable with an ion trap is substantially less than with an Orbitrap mass analyzer, thus leading to lower quality MS/MS spectra. This in turn may have affected *de novo* peptide sequencing, as on average, 8022 ± 4849 (s.d.) peptides were identified from each sample, with an average local confidence score of 30.5 ± 5.0 (s.d.), below MARLOWE's threshold score of 50. Comparatively, the average number of *de novo* peptides identified in the TCUP data set was 20,954 ± 8963 (s.d.), with an average local confidence score of 63.1 ± 5.8 (s.d.) (Supporting Figure S2). This disparity in characterization success suggests that MARLOWE requires high spectral data quality. While overall low taxonomic score distributions can arise from various reasons, such as taxonomic skew within the database, as in the case of *E. coli* in the TCUP section above, or low data quality in this case, the combination of low Novor scores and low taxonomic scores are a good indicator that the lists of returned organisms represent spurious hits rather than potential sources.

Similarly for the Danish peat bog data set, few ancient skin samples yielded successful characterization by MARLOWE. The modern skin samples from *Ovis aries* (sheep), *Capra hircus* (goat), and *Bos taurus* (cattle), on the other hand, yielded correct characterization at the species level despite high genetic similarity between sheep and goat, suggesting MARLOWE's ability to distinguish skin samples from domesticated animals. However, only one replicate of ancient skin samples found at the Haraldskaer site returned *C. hircus* as a potential contributor and three different skin samples (found in Mogelmose, Haraldskaer, and Huldemorose, respectively) returned *B. taurus* as potential contributors (Supporting Figure S3). The majority of the ancient skin samples returned ambiguous results from MARLOWE, whose organism lists do not include any of the potential contributors. Of the three ancient skin samples that yielded any characterization to potential contributors, only two align with the mass spectrometry-based characterization performed by Brandt et al.[28] Interestingly, the three potential contributors share the same family (Bovidae) and are assembled into the same taxonomic group. Even the modern skin samples contain tag-strong peptide matches to the other organisms within the same taxonomic group, though with substantially fewer tag-strong peptide matches than the correctly characterized organism. This observation indicates that the proteomes of goat, sheep, and cattle are very similar, which makes distinguishing among the three organisms more challenging. As such, it is not unexpected that ancient skin samples from these organisms, which have undergone degradation and may not be of the highest sample quality, add an extra layer of complexity to the analysis.

https://doi.org/10.1021/acs.jproteome.3c00477
J. Proteome Res. XXXX, XXX, XXX−XXX

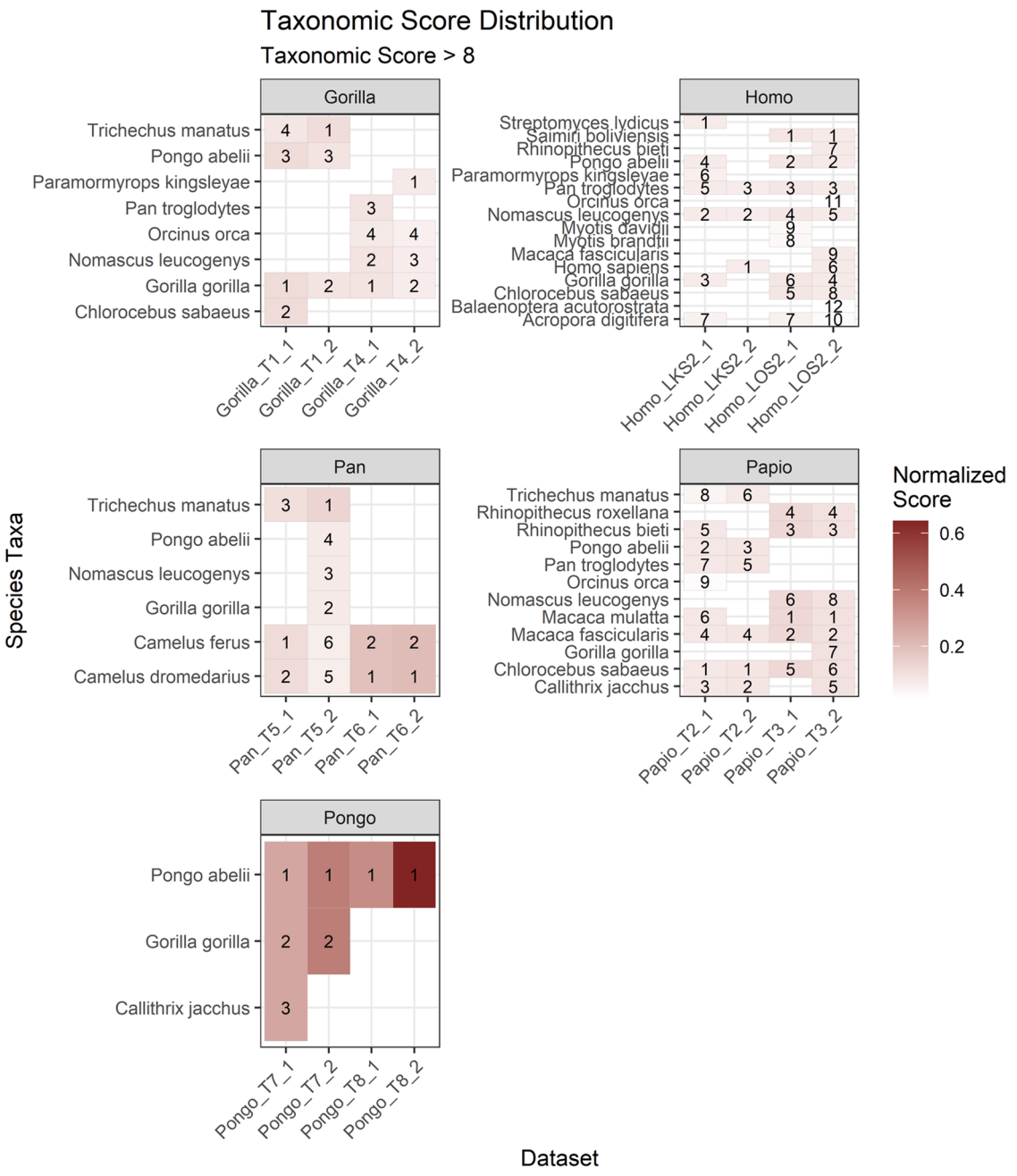


Figure 3. Heatmap of organism ranking of all taxa with taxonomic score >8 for each primate tooth sample, grouped by true contributor genera. Cells are colored by normalized taxonomic score, normalized to the total taxonomic score per sample. Values in each cell represent the organism rank, based on taxonomic score and the number of organism hits.

Results of these two data sets underscore the importance of sample and mass spectrometry data quality for MARLOWE's successful characterization. Low taxonomic score distributions within each sample represent a key indicator of low data quality and less-than-successful characterization.

Analysis of Mammalian Samples: Primate Teeth

We also examined MARLOWE's characterization specificity for mammalian samples and compare its performance to the specificity in microbial samples, i.e., the *B. cereus* data set[12] described in Jenson et al.[21] To that end, we analyzed a data set of tooth samples from closely related primates: human (*Homo*), gorilla (*Gorilla*), chimpanzee (*Pan*), orangutan (*Pongo*), and baboon (*Papio*), which were acquired by Froment et al.[27] Species identification from tooth samples is of forensic and archeological interest, as tooth samples often survive environmental insult and are a protein-rich sample type for source attribution and sex determination.[30,31] While we have utilized MARLOWE primarily for analysis of microbial samples, here, we expand our analysis to closely-related mammalian samples, since class-based forensic identification spans a broad range of taxonomies, of which the ability to characterize microbial samples and distinguish human from nonhuman samples are particularly important. In their analysis of these samples, Froment and co-workers fractionated proteins for each sample by molecular weight into those >10 kDa, and those between 3 and 10 kDa.[27] However, we focused our analysis on the first set of samples since we expect this group to contain the majority of proteins in the tooth samples,

https://doi.org/10.1021/acs.jproteome.3c00477
J. Proteome Res. XXXX, XXX, XXX−XXX

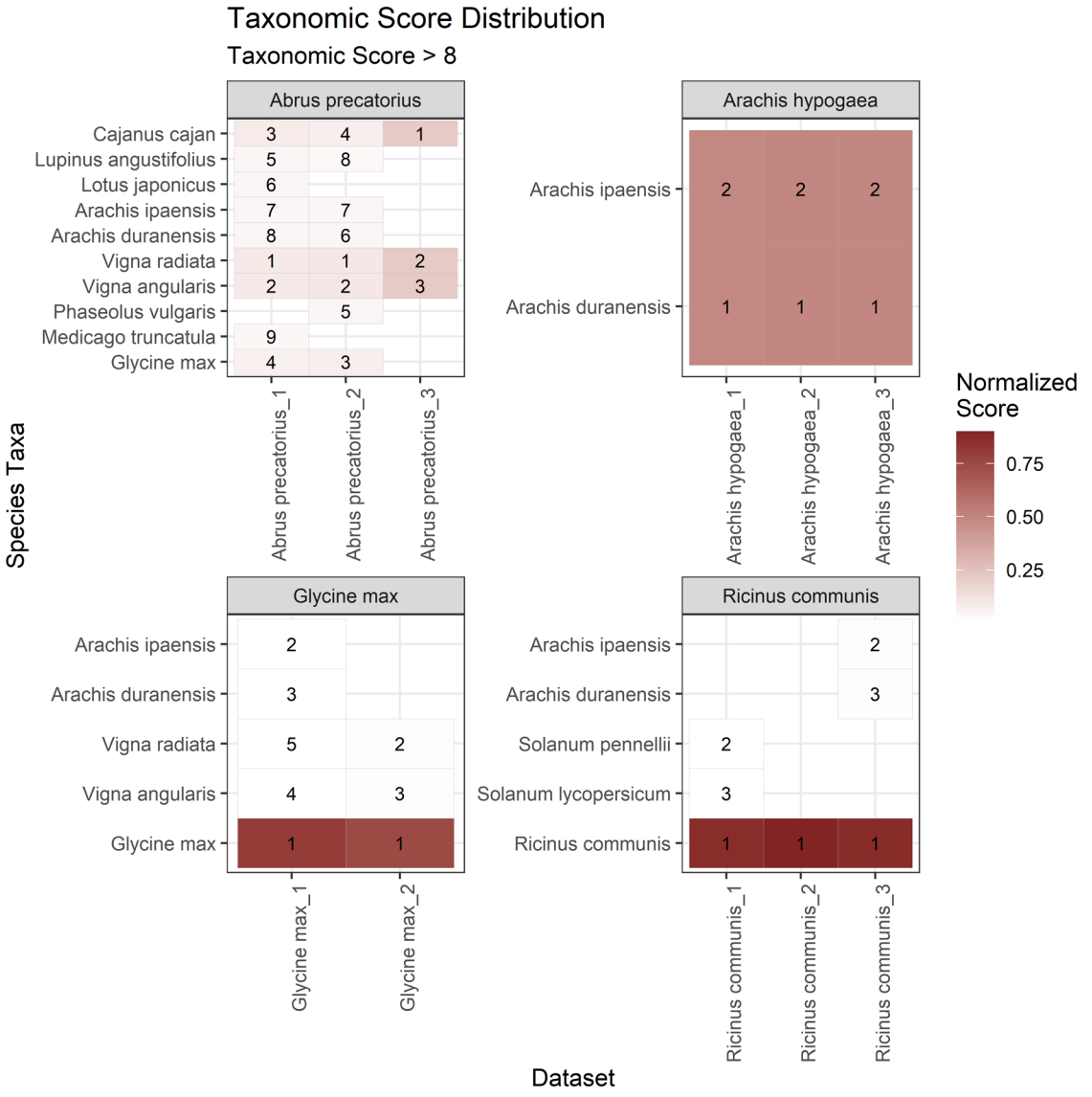


Figure 4. Heatmap of organism ranking of all taxa with taxonomic score >8 for each ground seed extract, grouped by true contributor species. Cells are colored by normalized taxonomic score, normalized to the total taxonomic score per sample. Values in each cell represent the organism rank, based on taxonomic score and the number of organism hits.

and thus, the most likely to yield correct taxonomic characterizations. For results including the latter set of samples, see Supporting Figure S4.

Analysis of this data set yielded correct characterization in 44% of samples for genera contained in the KEGG database, defined strictly as ranking the true contributor as the top hit. When considering ranked organisms with taxonomic score >8, MARLOWE returns the true contributor within the top six ranked organisms for 63% of samples (Figure 3). Of the five genera examined, only the *Papio* genus is not in the KEGG database, and so it is not surprising that the top-ranked organisms for *Papio* samples returned by MARLOWE are close relatives, that is, either *Macaca mulatta* (rhesus monkey) or *Chlorocebus sabaeus* (green monkey), both of which are in the same family as *Papio* (Cercopithecidae), and estimated to have diverged from baboons approximately 12 and 14 million years ago, respectively.[32] In comparison, the Hominoidea superfamily that consists of the *Homo*, *Gorilla*, and *Pongo* genera diverged from the Cercopithecoidea superfamily over 32 million years ago and are more distant relatives compared to the *Macaca* and *Chlorocebus* genera.[32]

Among the different groups of samples, correct characterization to the *Pongo* genus was most confident, with *Pongo abelii* as the top-ranked organism in all *Pongo* replicates and higher normalized taxonomic scores than other top-ranked organisms in other sample groups (on average, 0.41 ± 0.16 (s.d.) normalized score) (Figure 3). Correct characterization of the *Gorilla* genus as the top-ranked organism occurred in at least one of two replicates. Surprisingly, results from tooth samples of two individuals (Homo_LOS2 and Homo_LKS2), whose teeth were extracted 3 and 15 years ago, respectively, showed higher ranking of other primate species than the true contributor, including *P. abelii*, *Pan troglodytes*, and *Gorilla gorilla*, whose genera are of interest in this data set. In contrast, for all nonhuman primate teeth samples, *Homo sapiens* was not returned as a potential contributor even though other primates were listed and ranked (Figure 3). This result likely stems from the human proteome having the fewest strong peptides that enable distinction of this genus compared to other primates.

https://doi.org/10.1021/acs.jproteome.3c00477
J. Proteome Res. XXXX, XXX, XXX−XXX

Similarly, the chimpanzee proteome has the second fewest strong peptides of primates in the KEGG database, resulting in unsuccessful characterization of *Pan* tooth samples.

Compared to the results described in Jenson et al.[21] for the *B. cereus* data set, characterization of teeth samples of closely-related mammalian species appears to be more challenging for MARLOWE. Not only is the overall correct characterization rate for the data set lower (44 vs 88%), but the taxonomic scores for correct characterization to the top-ranked organism are also lower (18.7 ± 7.4 (m ± s.d.)), compared to those for the correctly characterized *B. cereus* samples (359.7 ± 250.4 (m ± s.d.)) (Jenson et al.)[21] Additionally, deeper examination of the taxonomic scores and organism ranking revealed that the four genera of interest (*Homo, Pongo, Pan, and Gorilla*) that are in MARLOWE's SQL database are in the same taxonomic group, and as such, have identical taxonomic scores. Thus, their distinct ranking within each sample is based solely on higher ranked organisms having a greater number of tag-strong peptide matches. For example, a *Pongo* tooth sample was correctly characterized to *P. abelii* as the top-ranked organism, but MARLOWE also returned *G. gorilla* as a potential contributor. The top two taxa had identical taxonomic scores of 21, but only 3 tag-strong peptide matches for *G. gorilla* compared to 18 matches for *P. abelii* (Figure 3). The few tag-strong peptide matches detected here may have manifested from limitations in protein diversity among primate tooth proteomes.

However, regardless of extent of primate teeth proteomic diversity, we note that organization of taxonomic groups for primates is substantially different from the taxonomic group assembly previously seen in the *B. cereus* data set, which has implications for proteomic diversity of these two groups. Where previously, multiple strains of *B. cereus* members could span multiple taxonomic groups, an indicator of proteome differences at the strain level, the four primate genera were contained in the same taxonomic group. This organization of taxonomic groups for primates in contrast to the *B. cereus* superspecies suggests that proteome differences among primates, and perhaps even other mammals, are much narrower than among microbes. In essence, the extent of proteome differences among primates at the family level may be small and similar to those among bacteria at the species level. Modern taxonomic nomenclature does not provide quantitative guidelines for taxon names. Given that MARLOWE relies on grouping organisms by their proteomic similarity (i.e., taxonomic group), it is not surprising that these primate teeth samples were correctly characterized at the family level, and correct characterization performance substantially improves at the taxonomic group level.

Characterization of Ground Seed Extracts: Seeds

Finally, we demonstrate MARLOWE's performance in characterizing ground seed extracts, an important forensic proteomics application as this sample type is commonly encountered in the identification of the protein toxin ricin.[1,2] This data set[33] contains biological replicates of *R. communis* (castor seed), *A. precatorius* (jequirity pea, the source of the toxin abrin), *G. max* (soybean), and *A. hypogaea* (peanut), prepared as described in Methods section above. MARLOWE successfully characterized all *R. communis* and *G. max* extracts as the top-ranked organism. Of the two true contributors not contained in the KEGG database, *A. hypogaea* and *A. precatorius*, the top-ranked organisms were close relatives, at the genus and subfamily clade levels,[34] respectively (Figure 4). These results demonstrate MARLOWE's success in characterizing seed extracts with high confidence, particularly in distinguishing castor seeds toward ricin identification in forensic proteomics.

Interpreting MARLOWE's Taxonomic Characterization

The intent of MARLOWE's development is to address a gap in forensic proteomics, to generate investigative leads on potential organisms in unknown biological samples without presumptions of source organisms in a forensic context. Here, we specifically focus on characterizing MARLOWE's performance on a number of forensically-relevant biological samples. Published organism identification/classification tools may report higher success rates than those reported here, but such studies do not typically explore the breadth of challenging samples/data sets addressed here (see below). The known lower accuracy of *de novo* identifications compared to database search may also play a role. In either case, understanding the limits of our tool was a key goal of this research.

As demonstrated by the different levels of MARLOWE's characterization success for each of the analyzed data sets, taxonomic source characterization highly depends on a combination of the following: high-quality, intact samples, mass spectral data quality, and taxonomic representation and underlying proteome relatedness of the organisms in the database. While high, distinct taxonomic scores for top-ranked organisms are simpler to interpret, low taxonomic scores can be more ambiguous and can arise from multiple reasons, including sample and data quality. Mass spectral data quality for *de novo* sequencing is crucial for success in MARLOWE, as observed from characterization of the Danish peat bog skin samples and extracts from plague papers. The combination of low taxonomic scores and low *de novo* peptide sequencing scores suggests suboptimal samples and/or data acquisition parameters.

Analysis of the other data sets showed differences in taxonomic characterization specificity, ranging from the capability to differentiate microbial samples at the species level and some resolution to different strains, to primates with lower characterization rates even at the family level. Though true contributors may not be the top-ranked hit, MARLOWE often returns true contributors in lists of potential organisms; this can be observed in the increased characterization rates when considering correct characterization in the top 5 ranks compared to the top 1 rank in Table 1. However, when reorganized into taxonomic groups based on proteomic similarity of the organisms in the SQL database, characterization rates to the true contributors as the top-ranked organism are similar or improve upon characterizing to assigned taxonomic levels (Table 1). Because taxonomic groups better represent proteome diversity, organisms in highly ranked taxonomic groups returned from MARLOWE can serve as investigative leads for further data analysis toward identification of unknown samples, such as guiding database searches with protein sequences from relevant organisms. In fact, we strongly encourage follow-up actions for confirmatory analyses, such as pairing database search with potential organisms suggested by MARLOWE, as the original intent of MARLOWE was not to produce organism identifications, but to narrow the list of potential organisms from all known organisms to a candidate list of statistically-likely organisms.

https://doi.org/10.1021/acs.jproteome.3c00477
J. Proteome Res. XXXX, XXX, XXX−XXX

We recognize that taxonomic representation in the organism reference database plays a large role in characterization success and can be another reason for observing low taxonomic scores, as in the case of *E. coli* samples. The KEGG Genome database that was used in MARLOWE includes many more prokaryotes than eukaryotic species, which somewhat skews the distributions of shared and strong peptides. Further, not all organisms are included in the KEGG Genome database, of which 5851 organisms are incorporated into MARLOWE, compared to the UniProtKB database which consists of 161,146 nonredundant organisms (accessed on September 2, 2021); this affects taxonomic characterization specificity. Notably, MARLOWE returns close relatives when proteomes of true contributors are not found in the SQL database, and lower taxonomic scores for top-ranked organisms can be an indicator that the true contributors are not in MARLOWE. We expect the skew in taxonomic representation and proteome relatedness would be moderated by replacing the KEGG Genome database with organisms in the UniProtKB database, which will improve taxonomic scores and characterization; this work is currently underway.

We selected the KEGG Genome database because its structure easily mapped peptides to organisms, and its size was not so great that it posed a significant database engineering challenge, allowing us to focus on other parts of the MARLOWE algorithm. The database includes ∼260 million tryptic peptides for the 5851 organisms present. Constructing a database to include all of the ∼1.3 million organisms present in UniProtKB is a major effort requires high-throughput computing resources. Thus, a comparison of MARLOWE's performance with different databases (e.g., KEGG Genome vs UniProtKB) that is outside the scope of this study. However, if a computationally-efficient analysis could be achieved with the larger UniProtKB database, comparison of its characterization performance with the current version of MARLOWE would be of value.

Though the focus of this work has been on forensically relevant samples, which we demonstrate with five analyzed data sets, this untargeted approach to metaproteomics analysis is not limited to forensic applications. Indeed, MARLOWE may find other utility in other bioanalytical and clinical applications to provide investigative leads of potential sources of unknown, complex mixtures in an unbiased and statistically robust manner.

■ ASSOCIATED CONTENT

Data Availability Statement

Publicly available raw mass spectrometry data were acquired from the following projects in ProteomeXchange and can be accessed via the online repository[24,25] (http://proteomecentral.proteomexchange.org) under accessions PXD004321, PXD018933, PXD008103, and PXD001029. The mass spectrometry proteomics data have been deposited to the PRIDE Archive (http://www.ebi.ac.uk/pride/archive/) via the PRIDE partner repository with the data set identifier PXD037607.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.3c00477.

MARLOWE parameters per data set (Table S1); heatmap of organisms returned by MARLOWE for the plague papers data set (Figure S1); comparison of *de novo* peptide local confidence score distributions between the plague papers and TCUP data sets (Figure S2); heatmap of organisms returned by MARLOWE for the Danish peat bog data set, grouped by the sample's potential contributor, as determined by mass spectrometry in Brandt et al. (Figure S3); heatmap of organism ranking for primate teeth samples (Figure S4) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Fanny Chu − *Chemical & Biological Signatures Group, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0000-0002-1114-6182; Phone: 509.372.4819; Email: fanny.chu@pnnl.gov

Authors

Sarah C. Jenson − *Chemical & Biological Signatures Group, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0000-0002-0807-5651

Anthony S. Barente − *Chemical & Biological Signatures Group, Pacific Northwest National Laboratory, Richland, Washington 99352, United States; Department of Genome Sciences, University of Washington, Seattle, Washington 98195, United States*

Natalie C. Heller − *Applied Statistics and Computational Modeling Group, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0000-0003-0088-067X

Eric D. Merkley − *Chemical & Biological Signatures Group, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0000-0002-5486-4723

Kristin H. Jarman − *Chemical & Biological Signatures Group, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* Present Address: K.H.J.: Karius Inc., 975 Island Dr., Suite 101, Redwood City, CA, 94065; orcid.org/0000-0003-4783-1277

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.3c00477

Author Contributions

[⊥]F.C. and S.C.J. contributed equally to this work. F.C.: Data analysis, manuscript writing. S.C.J.: Algorithm design and development, data analysis. A.S.B.: Algorithm development and testing. N.C.H.: Algorithm development. E.D.M.: Algorithm conception and design, manuscript writing. K.H.J.: Algorithm conception, design, and development.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Kristin D. Victry for preparing samples to generate the Seeds dataset. 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. PNNL-SA-173136.

REFERENCES

(1) 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*, 18−31.

(2) Heller, N. C.; Garrett, A. M.; Merkley, E. D.; Cendrowski, S. R.; Melville, A. M.; Arce, J. S.; Jenson, S. C.; Wahl, K. L.; Jarman, K. H. Probabilistic limit of detection for ricin identification using a shotgun proteomics assay. *Anal. Chem.* **2019**, *91* (19), 12399−12406.

(3) O'Bryon, I.; Tucker, A. E.; Kaiser, B. L. D.; Wahl, K. L.; Merkley, E. D. Constructing a Tandem Mass Spectral Library for Forensic Ricin Identification. *J. Proteome Res.* **2019**, *18* (11), 3926−3935.

(4) Kalb, S. R.; Barr, J. R. Mass spectrometric detection of ricin and its activity in food and clinical samples. *Anal. Chem.* **2009**, *81* (6), 2037−2042.

(5) Kalb, S. R.; Barr, J. R. Mass spectrometric identification and differentiation of botulinum neurotoxins through toxin proteomics. *Rev. Anal. Chem.* **2013**, *32* (3), 189−196.

(6) Kalb, S. R.; Goodnough, M. C.; Malizio, C. J.; Pirkle, J. L.; Barr, J. R. Detection of botulinum neurotoxin A in a spiked milk sample with subtype identification through toxin proteomics. *Anal. Chem.* **2005**, *77* (19), 6140−6146.

(7) Kaiser, B. L. D.; Hill, K. K.; Smith, T. J.; Williamson, C. H. D.; Keim, P.; Sahl, J. W.; Wahl, K. L. Proteomic analysis of four Clostridium botulinum strains identifies proteins that link biological responses to proteomic signatures. *PLoS One* **2018**, *13* (10), No. e0205586.

(8) 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.

(9) Kaiser, B. L. D.; Wunschel, D. S.; Sydor, M. A.; Warner, M. G.; Wahl, K. L.; Hutchison, J. R. Improved proteomic analysis following trichloroacetic acid extraction of Bacillus anthracis spore proteins. *J. Microbiol. Methods* **2015**, *118*, 18−24.

(10) Merkley, E. D.; Sego, L. H.; Lin, A.; Leiser, O. P.; Kaiser, B. L. D.; Adkins, J. N.; Keim, P. S.; Wagner, D. M.; Kreuzer, H. W. Protein abundances can distinguish between naturally-occurring and laboratory strains of Yersinia pestis, the causative agent of plague. *PLoS One* **2017**, *12* (8), No. e0183478.

(11) 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. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (2), 194−210.

(12) 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.

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

(14) Muth, T.; Behne, A.; Heyer, R.; Kohrs, F.; Benndorf, D.; Hoffmann, M.; Lehtevä, M.; Reichl, U.; Martens, L.; Rapp, E. The MetaProteomeAnalyzer: A Powerful Open-Source Software Suite for Metaproteomics Data Analysis and Interpretation. *J. Proteome Res.* **2015**, *14* (3), 1557−1565.

(15) Muth, T.; Kohrs, F.; Heyer, R.; Benndorf, D.; Rapp, E.; Reichl, U.; Martens, L.; Renard, B. Y. MPA Portable: A Stand-Alone Software Package for Analyzing Metaproteome Samples on the Go. *Anal. Chem.* **2018**, *90* (1), 685−689.

(16) Boulund, F.; Karlsson, R.; Gonzales-Siles, L.; Johnning, A.; Karami, N.; Al-Bayati, O.; Åhrén, C.; Moore, E. R. B.; Kristiansson, E. Typing and Characterization of Bacteria Using Bottom-up Tandem Mass Spectrometry Proteomics*. *Mol. Cell. Proteomics* **2017**, *16* (6), 1052−1063.

(17) Deshpande, S. V.; Jabbour, R. E.; Snyder, P. A.; Stanford, M.; Wick, C. H.; Zulich, A. W. ABOid: A software for automated identification and phyloproteomics classification of tandem mass spectrometric data. *J. Chromatogr. Sep. Tech.* **2011**, *S5* (2), No. 001.

(18) Alves, G.; Yu, Y.-K. Robust Accurate Identification and Biomass Estimates of Microorganisms via Tandem Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2020**, *31* (1), 85−102.

(19) 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. *Nat. Commun.* **2022**, *13* (1), No. 2458.

(20) 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.

(21) Jenson, S. C.; Chu, F.; Barente, A. S.; Crockett, D. L.; Lamar, N. C.; Merkley, E. D.; Jarman, K. H. MARLOWE: Taxonomic Characterization of Unknown Samples for Forensics Using De Novo Peptide Identification *bioRxiv* 2024.

(22) Johnson, R. S.; Searle, B. C.; Nunn, B. L.; Gilmore, J. M.; Phillips, M.; Amemiya, C. T.; Heck, M.; MacCoss, M. J. Assessing Protein Sequence Database Suitability Using < em > De Novo Sequencing *. *Mol. Cell. Proteomics* **2020**, *19* (1), 198−208.

(23) Wunschel, D. S.; Melville, A. M.; Ehrhardt, C. J.; Colburn, H. A.; Victry, K. D.; Antolick, K. C.; Wahl, J. H.; Wahl, K. L. Integration of gas chromatography mass spectrometry methods for differentiating ricin preparation methods. *Analyst* **2012**, *137* (9), 2077−2085.

(24) 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.

(25) 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.

(26) Ma, B. Novor: Real-Time Peptide de Novo Sequencing Software. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (11), 1885−1894.

(27) Froment, C.; Zanolli, C.; Hourset, M.; Mouton-Barbosa, E.; Moreira, A.; Burlet-Schiltz, O.; M?llereau, C. Protein sequence comparison of human and non-human primate tooth proteomes. *J. Proteomics* **2021**, *231*, No. 104045.

(28) Brandt, L. R.; Schmidt, A. L.; Mannering, U.; Sarret, M.; Kelstrup, C. D.; Olsen, J. V.; Cappellini, E. Species Identification of Archaeological Skin Objects from Danish Bogs: Comparison between Mass Spectrometry-Based Peptide Sequencing and Microscopy-Based Methods. *PLoS One* **2014**, *9* (9), No. e106875.

(29) D'Amato, A.; Zilberstein, G.; Zilberstein, S.; Compagnoni, B. L.; Righetti, P. G. Of mice and men: Traces of life in the death registries of the 1630 plague in Milano. *J. Proteomics* **2018**, *180*, 128−137.

(30) Wasinger, V. C.; Curnoe, D.; Bustamante, S.; Mendoza, R.; Shoocongdej, R.; Adler, L.; Baker, A.; Chintakanon, K.; Boel, C.; Tacon, P. S. C. Analysis of the preserved amino acid bias in peptide profiles of iron age teeth from a tropical environment enable sexing of individuals using amelogenin MRM. *Proteomics* **2019**, *19* (5), No. 1800341.

(31) Froment, C.; Hourset, M.; Sáenz-Oyhéréguy, N.; Mouton-Barbosa, E.; Willmann, C.; Zanolli, C.; Esclassan, R.; Donat, R.; Thèves, C.; Burlet-Schiltz, O.; M?llereau, C. Analysis of 5000 year-old human teeth using optimized large-scale and targeted proteomics approaches for detection of sex-specific peptides. *J. Proteomics* **2020**, *211*, No. 103548.

(32) Pozzi, L.; Hodgson, J. A.; Burrell, A. S.; Sterner, K. N.; Raaum, R. L.; Disotell, T. R. Primate phylogenetic relationships and

https://doi.org/10.1021/acs.jproteome.3c00477
J. Proteome Res. XXXX, XXX, XXX−XXX

divergence dates inferred from complete mitochondrial genomes. *Mol. Phylogenet. Evol.* **2014**, *75*, 165−183.
(33) 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.
(34) Lavin, M.; Herendeen, P. S.; Wojciechowski, M. F. Evolutionary Rates Analysis of Leguminosae Implicates a Rapid Diversification of Lineages during the Tertiary. *Syst. Biol.* **2005**, *54* (4), 575−594.

https://doi.org/10.1021/acs.jproteome.3c00477
J. Proteome Res. XXXX, XXX, XXX−XXX