Journal of proteome research

pubs.acs.org/jpr

Article

METATRYP v 2.0: Metaproteomic Least Common Ancestor Analysis for Taxonomic Inference Using Specialized Sequence Assemblies—Standalone Software and Web Servers for Marine Microorganisms and Coronaviruses

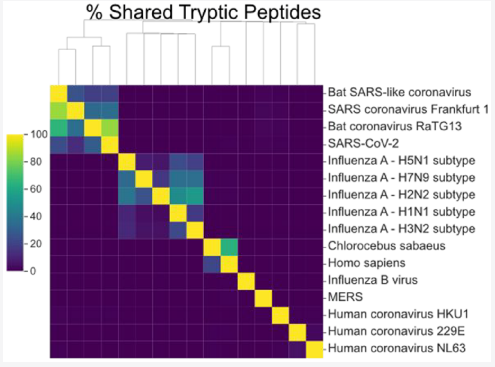
Jaclyn K. Saunders,* David A. Gaylord, Noelle A. Held, Nicholas Symmonds, Christopher L. Dupont, Adam Shepherd, Danie B. Kinkade, and Mak A. Saito*

Cite This: *J. Proteome Res.* 2020, 19, 4718–4729

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We present METATRYP version 2 software that identifies shared peptides across the predicted proteomes of organisms within environmental metaproteomics studies to enable accurate taxonomic attribution of peptides during protein inference. Improvements include ingestion of complex sequence assembly data categories (metagenomic and metatranscriptomic assemblies, single cell amplified genomes, and metagenome assembled genomes), prediction of the least common ancestor (LCA) for a peptide shared across multiple organisms, increased performance through updates to the backend architecture, and development of a web portal (https://metatryp.whoi.edu). Major expansion of the marine METATRYP database with predicted proteomes from environmental sequencing confirms a low occurrence of shared tryptic peptides among disparate marine microorganisms, implying tractability for targeted metaproteomics. METATRYP was designed to facilitate ocean metaproteomics and has been integrated into the Ocean Protein Portal (https://oceanproteinportal.org); however, it can be readily applied to other domains. We describe the rapid deployment of a coronavirus-specific web portal (https://metatryp-coronavirus.whoi.edu/) to aid in use of proteomics on coronavirus research during the ongoing pandemic. A coronavirus-focused METATRYP database identified potential SARS-CoV-2 peptide biomarkers and indicated very few shared tryptic peptides between SARS-CoV-2 and other disparate taxa analyzed, sharing <1% peptides with taxa outside of the betacoronavirus group, establishing that taxonomic specificity is achievable using tryptic peptide-based proteomic diagnostic approaches.



KEYWORDS: *metaproteomics, marine metaproteomics, marine microbiology, taxonomic annotation, tryptic peptide identification, coronavirus, COVID-19, SARS-CoV-2*

INTRODUCTION

In metaproteomics the mixture of a large number of organisms within each sample collected from a natural environment creates challenges in the attribution of peptides to specific proteins. This is especially problematic in instances where exact tryptic peptide sequences are shared between two or more organisms. This potential for shared peptides across proteins can create uncertainty in protein inference and taxonomic attribution. In bottom-up proteomics, the primary method used in metaproteomics to date, whole proteins are typically digested into smaller peptides with the enzyme trypsin. Since bottom-up metaproteomics directly measures these short tryptic peptides, as opposed to entire protein sequences, it is essential to understand the degree of shared peptides across proteins and taxonomic groups when assigning attributes of diverse environmental communities. Previously, we described the development of the METATRYP v 1 software which evaluates multiple organisms for shared peptides.[1] METATRYP v 1, METAproteomics of TRYPtic peptides, takes the full predicted proteome of an organism based on its reference genome, performs an in silico tryptic digestion of the proteins, and then stores the tryptic peptides of that organism within a single SQL database. Multiple taxa proteomes are stored within the SQL database. Using METATRYP tools, the database can be queried to identify how many taxa share a specific peptide sequence (or list of

Special Issue: Proteomics in Pandemic Disease

Received: June 1, 2020
Published: September 8, 2020

peptides), and it can also identify the total number of specific tryptic peptides shared across multiple organisms and for other phyloproteomic analyses. Previous applications of METATRYP v 1 using whole marine microbial genomes has aided in the development of targeted metaproteomic biomarkers for assessing environmental changes in space or time.[1−4] A useful result of the application of METATRYP v 2, with its extension of environmental data sequencing types, was the observation that the percentage of shared peptides between distinct marine microbial taxa was low, often in the single digit percentages, implying that the design of biomarker targets for species or even subspecies level analyses was tractable if sufficient care was taken.

In this manuscript, we describe version 2 of the METATRYP software (https://github.com/WHOIGit/metatryp-2.0). We have added additional features to improve its usability and performance. A major improvement was the addition of new data categories for different sequencing assembly methods, specifically those associated with de novo assembled metagenomic and metatranscriptomic data including metagenome assembled genomes (MAGs) as well as single cell amplified genomes (SAGs). METATRYP v 2 now supports these three specific data categories: "Genomes" for reference cultured isolates, "Specialized Assemblies" from SAGs and MAGs, and "Meta-omic assemblies" from metagenomic and metatranscriptomic assemblies. This greatly expands the utility of METATRYP since cultured genomes are often unavailable from natural environmental populations due to many organisms being difficult to culture with classical microbiological techniques,[5] or being only recently identified taxa. As a result, the availability of single cell genomes amplified and sequenced from the ocean environment, MAGs, and assembled metagenomics and metatranscriptomic data can contribute greatly to the identification and interpretation of metaproteomic data. Yet because these metagenomic and metatranscriptomic resources have varying levels of completeness and confidence in their functional and taxonomic assignments, maintaining them as separate categories of tryptic peptides within the database structure is particularly useful. In addition, METATRYP v 2 now supports the calculation of a least common ancestor (LCA) among shared tryptic peptides. For comparison, the Unipept web portal[6] has some similar functionality in identifying shared tryptic peptides and interpreting the Least Common Ancestor; however, Unipept relies on the Uniprot database which does not incorporate the wealth of environmental meta-omic sequencing available. Also, the Unipept portal does not support local curated database construction where users can evaluate unpublished sequencing resources like those of newly sequenced organismal genomes or novel environmental sequencing not yet available in the Uniprot curated database, whereas METATRYP can be installed locally for use with custom curated databases. The addition of these new sequence assembly data categories enables better prediction of shared peptides through enhanced representation of environmental sequence variability. The METATRYP marine web portal which is completely open for use by the research community currently contains a total of 182 354 079 unique peptides from 19 104 353 submitted protein sequences combined across all three data categories which will be expanded upon to include new marine-relevant sequences in the future. We welcome community requests for new sequences to be added to the online marine METATRYP database through an issue report on the GitHub repository (https://github.com/WHOIGit/metatryp-2.0/issues) or by email and Twitter.

Multiple improvements to the METATRYP software architecture and additional features were added to v 2. In order to improve performance speed and support these larger data categories (especially metagenomic and metatranscriptomic assemblies), the METATRYP SQL backend was converted from SQLite in METATRYP v 1 to a PostgreSQL backend in METATRYP v 2. Additional software and PostgreSQL implementations support the LCA analysis. METATRYP v 2 uses the same tryptic digest rules applied to METATRYP v 1, following trypsin-based digestion rules for proteins with peptides 6−22 amino acids in length.[1] Here we describe the technical improvements within METATRYP v 2, then demonstrate how metagenomics resources allow increased understanding of least common ancestor interpretations of metaproteomic results. Additionally, we provide an overview of the METATRYP web portal for marine microorganisms and the rapid deployment of a coronavirus-specific METATRYP web portal demonstrating the application of METATRYP to various research fields. Finally, a private API was added providing LCA analysis functionality to the Ocean Protein Portal,[7] enabling METATRYP to be inserted into other pipelines in the future.

IMPLEMENTATION

Database Backend Upgrades

METATRYP is built upon a Relational Database Management System (RDMS). Version 1 was built using a SQLite database. While this database management system was sufficient for single reference genomes, it was lacking in the speed needed for expanded sequencing data categories. In order to increase the speed of database construction (ingestion of proteomes), database searching, and expanded functionality, we have upgraded the database management system to a PostgreSQL backend which is an object-relational database. The Postgres backend has provided improvements in speed, as well as enhanced flexibility in searching.

Support for New Data Categories

The advent of environmental de novo nucleotide sequencing and assembly has identified an entire realm of microorganisms previously unknown to the scientific world, as many environmental microorganisms are not readily isolated using classical microbiological techniques.[5] METATRYP v 1 focused on the construction of a search database using reference organismal genomes, METATRYP v 2 is capable of handling newer types of sequencing and assemblies thus opening the search space to a much greater range of organisms likely found within the environment of interest. Within the field of metaproteomics, there has been great emphasis placed on the need for curated and appropriate search databases to be utilized for completing accurate peptide-to-spectrum matching (PSM) and downstream biological inference and interpretation;[8,9] this also holds true for the construction of a METATRYP database for evaluation of shared tryptic peptides in an environmental sample. METATRYP relies upon protein sequences predicted from genomic sequencing and does not currently take into account spectral data directly or any post-translational modifications (PTMs) to peptides. An analysis using METATRYP can be utilized to identify potential peptide targets for quantitative proteomics before analysis of a sample on a mass spectrometer, or it can be done in a post hoc fashion

https://dx.doi.org/10.1021/acs.jproteome.0c00385
J. Proteome Res. 2020, 19, 4718−4729

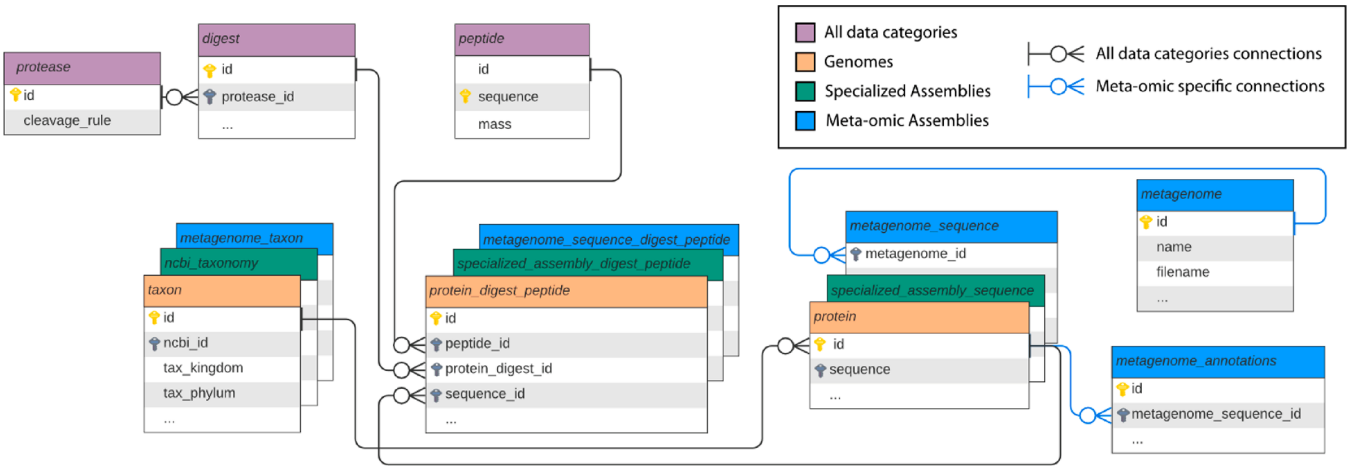


Figure 1. Core elements of the METATRYP v 2 database schema. An entity relationship diagram depicting the core tables for the three sequencing categories (orange tables: "Genome", green tables: "Specialized Assembly", and blue tables: "Meta-omic Assembly" data categories). The purple tables are shared tables among all three data categories containing information about the tryptic digestion rules (*protease* and *digest* tables) as well as the amalgamation of all unique tryptic peptide sequences found across all three data categories (*peptide*). The lines connecting the tables represent links between the data tables. The three data categories are stacked where tables represent similar information for each category. The metagenome data category requires two additional data tables as there are multiple taxa stored within a single meta-omic assembly sequencing file which requires an additional *metagenome_annotations* table for parsing; the blue connecting lines represent meta-omic specific data linkages.

to analyze taxonomic specificity on a global proteomics analysis after PSM identification with an appropriate sequencing database.[9]

In order to expand the environmentally relevant search space, METATRYP v 2 now handles newer assemblies from sources like metagenome and metatranscriptome assemblies, metagenome-assembled-genomes (MAGs), and single cell amplified genomes (SAGs). The incorporation of these newer data categories, in addition to the traditional single organism reference genome, greatly expands the environmental variability (and therefore, potential for shared tryptic peptides) within an environmental sample. In order to manage the larger sequencing databases generated by incorporation of this environmental data, and thus the greater burden of a larger search space, improvements were made to the back-end search database (see section Database Backend Upgrades). The database schema (Figure 1) for METATRYP was expanded not only to include these different data categories, but also to identify them as separate search spaces as the uncertainty of taxonomic identity among the sequencing categories is variable and should be taken into consideration during interpretation. The new data categories roughly mirror the original reference genome schema. However, additional tables are required to properly map "Meta-omic" data as multiple taxa are contained within a single "Meta-omic" assembly file.

Environmental sequencing data is also more likely to have more frequent occurrences of ambiguous bases in assemblies. These are base locations where it is uncertain what the correct amino acid should be, sometimes a result of low-quality base calling by the sequencing technology or due to a single location where there are multiple possibilities for the amino acid at that single location in the assembly that cannot be determined. METATRYP v 2 will recognize ambiguous bases, specifically the base symbol "X", which represents the presence of an unknown or ambiguous amino acid during the ingestion phase. As the specific amino acid represented by "X" is unknown, the exact tryptic peptide cannot be predicted. During ingestion, METATRYP will identify proteins which contain an "X" and report the affected protein to the standard output (stdout) stream. If there is a tryptic peptide containing the "X", that peptide will not be included in the METATRYP peptide table; however, all other tryptic peptides from that protein will be included in the peptide table.

Least Common Ancestor Analysis

Sequence homology is conserved among more closely related organisms. However, it is possible that tryptic peptides, especially shorter ones, may occur by chance or horizontal gene transfer across multiple taxa without a direct shared ancestry. Shared tryptic peptides have been previously shown to occur among disparate taxa in the marine environment.[10,11] In order to identify the occurrence of shared tryptic peptides either through shared evolutionary history or through variance in sequence as a result of a novel horizontal gene transfer event or sequence stochasticity, we have added least common ancestor (LCA) analysis to METATRYP v 2. The LCA analysis incorporates the phylogenetic lineage of the sequences imported into the METATRYP databases (Figure 1), then calculates the "least common ancestor" by finding the unifying phylogenetic point for all the organisms containing the tryptic peptide queried. In order for METATRYP to identify the common point in the taxonomic lineage, it requires a consistent taxonomic lineage to be used across the database for each proteome submitted. For METATRYP the shared phylogeny of the taxonomic groups is identified by pulling the taxonomic lineages from the National Center for Biotechnology Information (NCBI) Taxonomy Database.[12] For the creation of a user-generated METATRYP database capable of LCA analysis, the user can submit the NCBI taxon id number (taxid) for the input sequence files, and METATRYP will pull the taxonomic lineage information for each organism using Biopython[13] and Pandas[14] libraries in Python 3. This lineage information is then used to calculate the LCA among the organisms with shared peptides via the PostgreSQL Longest Common Ancestor function, which METATRYP uses to return the LCA for each sequencing data category.

https://dx.doi.org/10.1021/acs.jproteome.0c00385
J. Proteome Res. 2020, 19, 4718−4729

Genomic and Metagenomic Results

Peptide	Genome LCA	Metagenome LCA	Genomes Returned	Metagenomes Returned
LSHQAIAEAIGSTR	Cyanobacteria	Unclassified_Cyanobacteria	48	49
VNSVIDAIAEAAK	Prochlorococcus	Prochlorococcus	10	8

Genome Name	NCBI Id	IMG Id
Prochlorococcus sp. MIT9202	93058	647533199
Prochlorococcus sp. MIT9215	93060	640753041
Prochlorococcus sp. MIT9301	167546	640069322
Prochlorococcus sp. MIT9302	74545	2606217691
Prochlorococcus sp. MIT9311	167547	2606217680
Prochlorococcus marinus MIT9312	74546	637000210
Prochlorococcus sp. MIT9314	167548	2606217312
Prochlorococcus sp. MIT9321	167549	2606217683
Prochlorococcus sp. MIT9322	167550	2606217679
Prochlorococcus sp. MIT9401	167551	2606217316

Metagenome	NCBI ID	Taxon	Taxon Level
GOS/OMZ (JCVI)	1218	Prochlorococcus	genus
GOS/OMZ (JCVI)	1218	Prochlorococcus	genus
GOS/OMZ (JCVI)	1218	Prochlorococcus	genus
GOS/OMZ (JCVI)	1218	Prochlorococcus	genus
GOS/OMZ (JCVI)	1218	Prochlorococcus	genus
GOS/OMZ (JCVI)	1218	Prochlorococcus	genus
METZYME PEPTIDES (JGI/JCVI)	93058	Prochlorococcus marinus str. MIT 9202	species
METZYME FUSION PEPTIDES (JGI/JCVI)	93058	Prochlorococcus marinus str. MIT 9202	species

VAAEAVLSMTK	Synechococcales	Synechococcales	18	32

Figure 2. "Genomic" and "Meta-omic" query results from the marine METATRYP web portal for the peptides LSHQAIAEAIGSTR, VNSVIDAIAEAAK, VAAEAVLSMTK. The LCA results for these three separate peptides indicate the varying degrees of taxonomic uniqueness among the peptides in the "Genome" and "Meta-omic Assembly" data categories. In these two data categories, LSHQAIAEAIGSTR is unique to the Phylum Cyanobacteria. VNSVIDAIAEAAK is unique to the genus *Prochlorococcus*. VAAEAVLSMTK is unique to the Order Synechococcales. All of these peptides show potential as biomarkers at these varying taxonomic levels according to evaluation by the "Genome" and "Meta-omic Assembly" databases.

Web Portal and API

A primary goal of releasing the METATRYP software originally was to enable other users to create and curate customized databases for searching tryptic peptides, specifically with a focus on marine microbial communities. Prior metaproteomic analyses have identified the importance of curated search databases for peptide-to-spectrum matching, recommending the use of custom metagenomic/metatranscriptomic libraries tailored to the metaproteomic sample. Often, the nucleotide-based libraries used are those collected from the same environmental location as the proteomics sample. This maximizes the identification of high quality spectra while minimizing search space which can have detrimental impacts on speed and false discovery rates.[8,15] METATRYP v 2 expands on this goal through the creation of a web server which can be queried easily by users, without the need to install and run the software locally. The METATRYP v 2 site can be found at https://metatryp.whoi.edu/. This web server takes as input a peptide sequence, multiple peptides, or a full protein sequence submitted into a text box by a user which is then in silico digested into tryptic peptides. METATRYP then searches for the occurrence of these peptides across three different marine-specific data categories: an organismal reference genome data category ("Genomes") same as METATRYP v 1 (SI Table S1), and new data categories for "Specialized Assemblies" (SI Table S2) which currently contain 4783 Archaeal and Bacterial MAGs[16,17] assembled by binning metagenomic sequences[18−20] from the TARA Oceans sequencing project,[21] and a metagenomic and metatranscriptomic assembly data category ("Meta-omic" assembly)[3,4,22] (SI Table S3) which currently contains 4 863 985 predicted proteins. Ideally, additional MAGs and SAGs will be added to the METATRYP web portal database in an effort to broaden taxonomic coverage in the marine environment. The addition of Eukaryotic SAGs[23] would significantly extend the diversity of the current database.

Results from a METATRYP query are then returned to the user in an interactive drop-down table, showing the presence of the peptides within these data categories and the LCA result for each category (Figure 2). This example shows the results for three different peptides that have been searched simultaneously (LSHQAIAEAIGSTR, VNSVIDAIAEAAK, VAAEAVLSMTK) which are used for the default search on the web portal if a user does not enter a sequence query. To expand the search results for each peptide, the user clicks on the peptide sequence link; shown here are the results for peptide VNSVIDAIAEAAK in the "Genome" and "Meta-omic" data categories. Displayed are the taxa and their associated NCBI Taxonomy ID numbers (taxid) which link out to that taxon's entry within the NCBI Taxonomy Database. For the "Genome" category, the Joint Genome Institute (JGI) Integrated Microbial Genomes and Microbiomes (IMG) genome IDs are also shown, where available, with links out to IMG as the predicted proteomes for all those in the "Genome" category were collected from in the current version of this database.[24]

METATRYP can also compare entire organismal proteomes to identify the frequency of shared peptides across taxa within a given sequencing data category. This feature previously existed in METATRYP v 1 for generating peptide redundancy tables within the "Genome" sequencing data category for the selected full marine microbial genomes in the database and was used to show the relatively low occurrence of shared peptides across disparate taxa (Genus level and above classification

https://dx.doi.org/10.1021/acs.jproteome.0c00385
J. Proteome Res. 2020, 19, 4718−4729

levels) in the open ocean microbiome.[1] This feature can now be implemented within METATRYP v 2 on all three major data categories: "Genome", "Meta-omic Assemblies", and "Specialized Assemblies". Within the web portal, there is now a visualization tool for creating ordered heatmaps of shared peptide frequencies among taxa for the "Genome" and "Specialized Assembly" data categories. This visualization page, "Peptide Redundancy Heatmaps", was built in Python 3.7 using the Jupyter environment,[25] Pandas,[14] and Seaborn.[26] Users can select what taxa they wish to compare within a given data category, and a heatmap identifying the percentage of shared peptides compared to total peptides among the selected taxa is generated and displayed on the page below in a .png format. Displayed in the heatmaps are the percentage of pairwise shared peptides between taxa in a specified data category where the percentage is calculated as the number of shared peptides between taxon A and taxon B divided by the total number of peptides in taxon A. Given the varying levels of genome completeness for a specific taxon in the "Specialized Assembly" data category, this percentage should be viewed with more caution. Due to the aggregate nature of meta-omic assemblies, where many taxa of highly variable coverage depth are present within each data set, this heatmap visualization feature is not currently supported in the web portal for this data category.

Example of Rapid Deployment: Coronavirus Domain Application

The capabilities of the METATRYP software make it applicable to scientific domains outside of the marine microbial ecology and biogeochemistry fields. The ability to identify shared peptides in metaproteomics is critical to other metaproteomics studies of mixed communities, especially in the development of biomarkers for targeted metaproteomics. An example application of METATRYP v 2 to other domains is the creation of a coronavirus-focused database and the rapid deployment of a coronavirus-focused METATRYP web portal (https://metatryp-coronavirus.whoi.edu). The database for this web portal uses the predicted proteomes for multiple riboviruses with a focus on betacoronaviruses, including SARS-CoV-2 responsible for the COVID-19 outbreak,[27] SARS-CoV-1 strains like SARS strain Frankfurt 1[28] isolated during the 2003−2004 SARS outbreak, Middle East Respiratory Syndrome-related (MERS) coronavirus strains,[29] and strains associated with the common cold[30] like human coronavirus strains NL63,[31] HKU1,[32] and 229E[33] for a total of 94 coronavirus taxa in the database. It also contains the human proteome, the African Green Monkey (*Chlorocebus aethiops sabaeus*) proteome as it is the taxonomic source of the Vero cell line commonly used in virus replication studies and plaque assays,[34] common oral bacteria,[35] six *Lactobacillus* strains associated with the human microbiome,[36] the most common influenza strains (Influenza A: H1N1 and H3N2; Influenza B) as well as other influenza strains, and common proteomic contaminants in the common Repository of Adventitious Proteins (cRAP). All taxa included in the coronavirus database and their associated NCBI Taxonomy IDs (SI Table S4) are listed on the databases page in the web portal (https://metatryp-coronavirus.whoi.edu/database). In order to capture the variability of sequences, we downloaded all the proteins (aside from those in the cRAP database) for each taxon from the NCBI Identical Protein Groups (IPG) Database using taxon sequence identifiers (SI Table S4). The IPG Database enables collection of a single nonredundant entry for each protein translation found from several sources at NCBI, including annotated coding regions in GenBank and RefSeq, as well as records from SwissProt and PDB.[37] One sequence for each identical protein group was collected for NCBI taxa with >10 identical protein groups, collecting proteins from the specified taxon and all its children from the NCBI Taxonomy Database. However, for the taxon "Severe acute respiratory syndrome-related coronavirus" (NCBI taxid: 694009), only protein groups from that specific taxid level were recruited (using the flag "txid694009[Organism:noexp]" in the query). This taxon should only contain SARS-CoV-1 related proteins; however, it may contain some non-SARS-CoV-1 sequences due to inconsistent nomenclature during the emergence of this relatively new pathogen. Proteins were collected using Biopython[13] and the NCBI Entrez[38] E-Utilities API.[39]

By using the Identical Protein Groups, the database captures sequence variability while reducing redundancy in database construction. This nonredundant collection of protein sequences per taxon is essentially the collection of the known pan-proteomes for each taxon; it is the representation of the predicted proteome of a single organism's genome plus all the sequence variation captured from sampling a population of organisms within a taxonomic group. For example, the taxon *Homo sapiens* pan-proteome from IPG has >1 000 000 proteins capturing sequence variability from the population of human sequences in the NCBI database, whereas an individual human genome contains <20 000 protein coding genes.[40] Due to the varying sequencing efforts among some taxa, the length of the proteomes in this database may vary. For example, the taxon severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, 2019-nCov, COVID-19 virus; taxid 2697049) has 1752 unique Identical Protein Groups, whereas the taxon Bat coronavirus isolate RaTG13 (taxid 2709072), SARS-CoV-2's closest sequenced animal virus precursor,[41] has 10 unique Identical Protein Groups as of the download date (May 3, 2020) even though the genomes of these viruses are roughly the same length. Since the population of SARS-CoV-2 has been sequenced more frequently, more sequence variability has been captured in the NCBI databases. Due to the collection of the pan-proteomes for the coronavirus METATRYP database, the peptide redundancy heatmaps need to be viewed with caution, as taxa which have received a higher degree of sampling effort will have more peptides associated with them. Therefore, it is important to take into consideration both combinations of pairwise taxa comparisons for heatmap calculations (both sides of the heatmap separated by the diagonal). For example, one should evaluate the percentage of shared peptides between SARS-CoV-2 and RaTG13 where the total number of tryptic peptides for SARS-CoV-2 is in the denominator and also where the total number of tryptic peptides in the denominator is for RaTG13. When calculating peptide redundancy between organisms, METATRYP reports this calculation as "individual percent".

$$\text{individual percent} = \left(\frac{\text{peptides in taxon } A \cap \text{peptides in taxon } B}{\text{peptides in taxon } A}\right) \times 100\%$$

The percentage of shared peptides across taxa can also be calculated by the number of peptides shared between taxa with the combined total number of peptides for both taxa in the denominator, METATRYP reports this as "union percent".

https://dx.doi.org/10.1021/acs.jproteome.0c00385
J. Proteome Res. 2020, 19, 4718−4729

Specialized Assembly Results

Peptide	Lowest Common Ancestor	MAGS Returned
VNSVIDAIAEAAK	Prochlorococcus	34
VAAEAVLSMTK	Synechococcales	42
LSHQAIAEAIGSTR	Bacteria	95

Genome Name	Species	NCBI ID	Study Name
TARA_ASW_MAG_00003	Unclassified Cyanothece	43988	Delmont_TARA_MAGs
Prochlorococcus_marinus_SCGC_AAA795-M23	Prochlorococcus marinus	1219	Tully
Prochlorococcus_marinus_SCGC_AAA795-I06	Prochlorococcus marinus	1219	Tully
TARA_ION_MAG_00012	Prochlorococcus marinus	1219	Delmont_TARA_MAGs
TARA_ANW_MAG_00068	Unclassified Cyanobium	167375	Delmont_TARA_MAGs
...	...	...	...
Cyanobacteria_bacterium_UBA6047	Cyanobacteria bacterium UBA6047	1947888	Tully
Verrucomicrobiales_bacterium_strain_NP1000	Verrucomicrobiales bacterium	2026801	Tully
Synechococcus_sp_WH7805	Synechococcus sp. WH 7805	59931	Tully

Figure 3. "Specialized Assembly" abbreviated query results from the marine METATRYP web portal for the peptides LSHQAIAEAIGSTR, VNSVIDAIAEAAK, and VAAEAVLSMTK. Among the MAGs currently in the METATRYP web portal database, VNSVIDAIAEAAK and VAAEAVLSMTK are unique to MAGs at the Genus level for *Prochlorococcus* and *Synechococcus*, respectively. However, the peptide LSHQAIAEAIGSTR, while indicating specificity at the Phylum level of Cyanobacteria in the "Genome" and "Meta-omic Assembly" data categories, reports 94 Cyanobacterial MAGs with this peptide and 1 Verrumicrobia MAG with this peptide, warranting further investigation of this peptide as a potential Cyanobacterial biomarker and considerations for use of this potential biomarker in environments which may have abundant Verrumicrobial populations.

$$\text{union percent} = \left(\frac{\text{peptides in taxon } A \cap \text{peptides in taxon } B}{\text{peptides in taxon } A \cup \text{peptides in taxon } B}\right) \times 100\%$$

However, this also needs to be interpreted with care as when comparing a taxon that may have only been sequenced once (few total peptides) with a broadly sampled taxon (many peptides due to environmental variability), the signal of the rarely sampled taxon may be reduced due to the large n of the heavily sequenced organism. In addition, viewing "union percents" when comparing an organism with a small proteome vs an organism with a large proteome would also skew any signal of shared peptides; for example, SARS-CoV-2 only encodes 12 proteins/genome,[27] whereas the human genome encodes ~120 000 proteins. Even though the NCBI IPG Database is not used in the marine METATRYP web portal, this same effect may be observed when comparing organisms with highly uneven proteome sizes, say if marine phages are added to the database, or with taxa that have varying levels of genome sequencing completeness, such as with "Specialized Assembly" data like MAGs and SAGS.

RESULTS AND DISCUSSION

Backend Upgrades for METATRYP v 2 Provide Improved Performance

The switch from a SQLite database backend in METATRYP v 1 to a PostgreSQL database backend has resulted in significant improvements in performance and functionality of METATRYP for the research community. In particular, this transition has resulted in improved computational times and facilitated the addition of LCA analyses to the software package. To test computational speed, a database was constructed based upon the reference genomes from 136 marine microbial taxa (SI Table S5) in both versions of METATRYP. SI Table S6 shows the benchmarks associated with the construction of these genome-only databases for comparison, with METATRYP v 2.0 taking only 17% of the time it took METATRYP v 1.1 to construct the same database. By converting to a PostgreSQL backend, the computational time to query the example database with the example peptide "LSHQAIAEAIGSTR" for an exact match was reduced by 4×, dropping from 0.41 s using METATRYP v 1.1 to 0.11 s with METATRYP v 2.0. The CPU utilization for METATRYP v 2.0 was also lower than for v 1.1.

It is noted that setting up a PostgreSQL server on a local machine requires administrative permissions and is more complex than SQLite which is more user-friendly and requires fewer system dependencies. Therefore, for users without local administrative permissions, METATRYP v 1 with its SQLite backend remains a good lighter weight option (albeit with slower performance and without the capacity to handle diverse data categories and LCA analyses). The PostgreSQL backend facilitates the LCA analysis through a call to the PostgreSQL Longest Common Ancestor function. In order to facilitate ease of setup for the marine microbial research community, we have added a copy of this preconstructed marine microbial METATRYP v 1 SQLite database to the METATRYP v 1 code repository (https://github.com/saitomics/metatryp). A complete preconstructed marine microbial database for METATRYP v 2 containing sequence data for 136 genomes (SI Table S5) and the base schema for all three data categories is included in the METATRYP v 2 code repository (https://github.com/WHOIGit/metatryp-2.0).

Use of New Data Categories in METATRYP for Marine Microbial Communities

Using the results of the peptides set as the default example search in the METATRYP web portal (LSHQAIAEAIGSTR, VNSVIDAIAEAAK, VAAEAVLSMTK), we see the results for these three peptides have varying levels of taxonomic LCAs, ranging from Species- to Phylum-levels of taxonomic specificity (Figures 2 and 3). For peptide VNSVIDAIAEAAK, the LCA across all data categories is the genus *Prochlorococcus*, indicating that this peptide appears unique to this genus in the marine microbial community and is therefore a potential biomarker for targeted metaproteomics that allows species level specificity. For peptide VAAEAVLSMTK, the LCA for all categories is the Order Synechococcales as this peptide is

https://dx.doi.org/10.1021/acs.jproteome.0c00385
J. Proteome Res. 2020, 19, 4718−4729

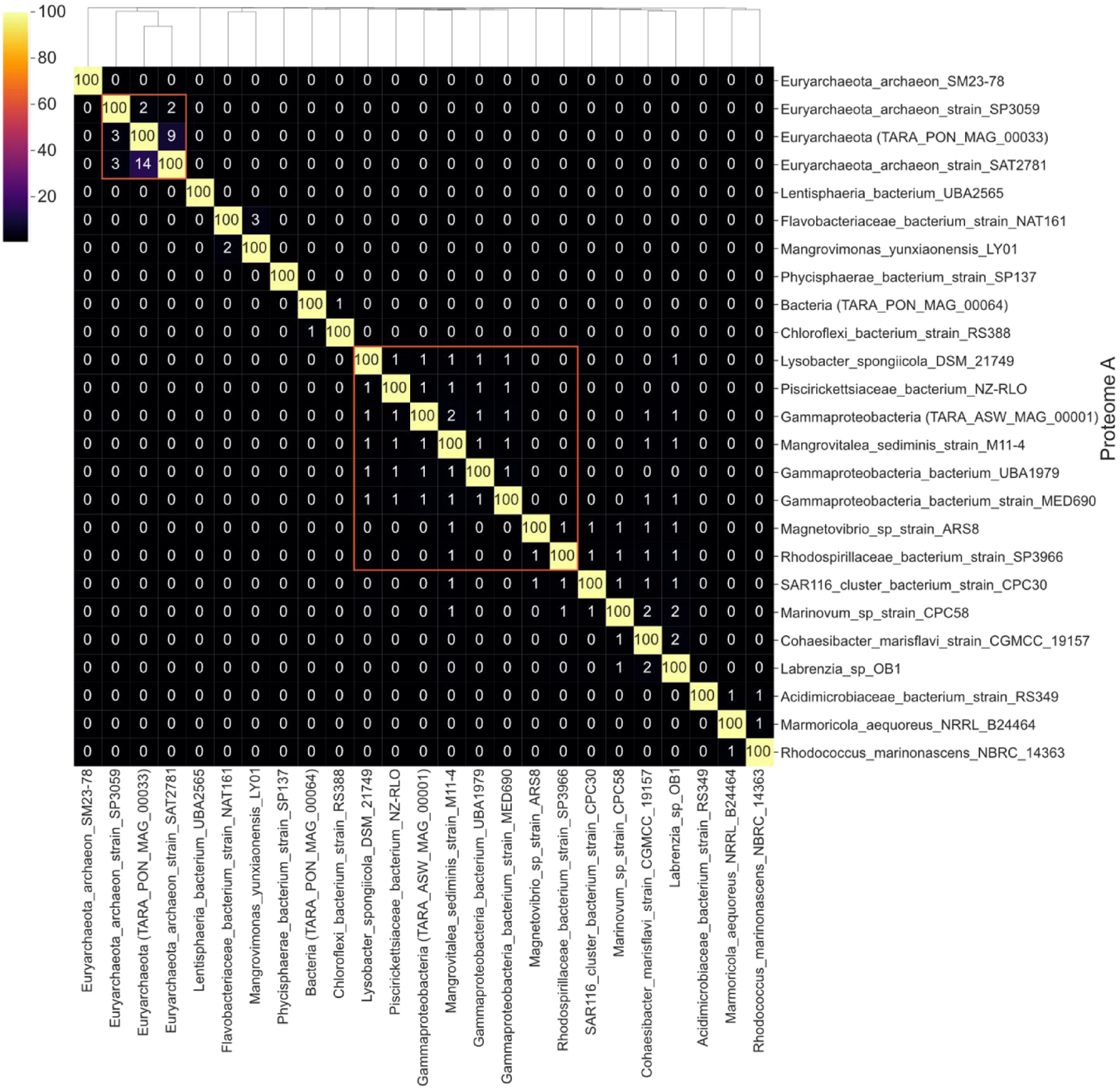


Figure 4. Heatmap with cluster dendrogram displaying the individual percents of shared tryptic peptides across 25 randomly selected MAGs from the new "Specialized Assembly" data category. The individual percent is calculated as the percentage of shared peptides across Proteome A and Proteome B divided by the number of total peptides in Proteome A. The data here show a very low frequency of shared tryptic peptides across this random sample of MAGs (SI Figures S2 and S3 show 100 randomly selected MAGs displaying a similar trend). The red outlines highlight two example regions of Euryarchaeotal and Gammaproteobacterial MAGs showing some overlap in shared tryptic peptides among these taxonomic groups.

found within the Genus *Prochlorococcus* and its sister Genus *Synechococcus*.

Taxonomic assignment of the original sequences of predicted proteins for the different data categories ranges in uncertainty. From reference genomes from cultured isolates being the most certain, to metatranscriptomic and metagenomic assemblies providing more high-scoring PSMs, but also having the least certainty in taxonomic assignment of the source proteins. The "Specialized Assemblies" of SAGs and MAGs exist somewhere in between on this taxonomic assignment uncertainty spectrum. Peptide LSHQAIAEAIGSTR (Figure 3) demonstrates this level of uncertainty within the "Meta-omic" assembly category as a source protein for this peptide in the GOS/OMZ (JCVI) metagenome cannot be identified below the Phylum level of Cyanobacteria (Figure 2) and in the "Specialized Assembly" category containing thousands of MAGs, this peptide is found in 94 Cyanobacterial MAGs and one Verrumicrobial MAG (Verrucomicrobiales_bacterium_strain_NP1000), resulting in a LCA of "Bacteria". Notably, this peptide is from the Global Nitrogen Transcriptional Regulation Protein (NtcA) which is highly conserved across the Cyanobacteria.[42] While this peptide has been

https://dx.doi.org/10.1021/acs.jproteome.0c00385
J. Proteome Res. 2020, 19, 4718−4729

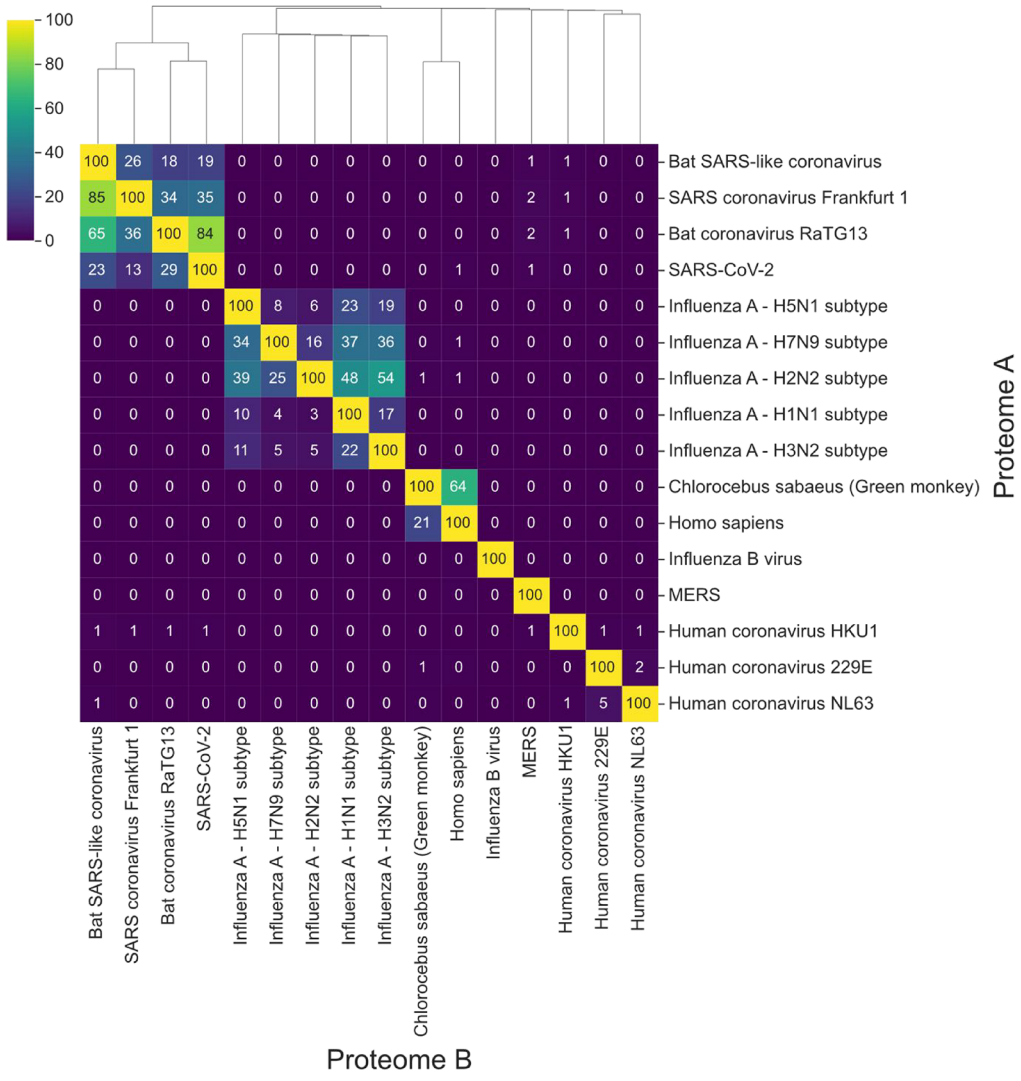


Figure 5. Heatmap with cluster dendrogram displaying shared tryptic peptides among a subsample of the taxa in the METATRYP coronavirus web portal database. In general, there is a very low occurrence of shared tryptic peptides (<1%) across disparate taxa. Higher frequencies of shared tryptic peptides are shown among more closely related taxonomic groups. Severe acute respiratory syndrome-related coronaviruses form a cluster in the top left corner. MERS-related coronavirus shares <1% of tryptic peptides with all taxa depicted here. The "common cold" strains of coronavirus (HKU1, 229E, and NL63) form a separate cluster in the bottom right corner. These coronavirus clusters are distinct and very different from the influenza A cluster in the middles, as the severe acute respiratory syndrome-related viruses share <1% of shared tryptic peptides with influenza A and B. *Homo sapiens* and *Chlorcebus sabaeus* form a distinct group, sharing more tryptic peptides with each other than with any other taxonomic groups.

previously identified as a potential biomarker for Cyanobacteria,[1] the addition of the "Specialized Assembly" data category identified the possible presence of this peptide in another phyla warranting further investigation. Interestingly, upon further investigation, the protein in MAG Verrucomicrobiales_bacterium_strain_NP1000, identified from metagenomes in the Red Sea,[17] is >99% identical with an NtcA in the Cyanobacterium genus *Synechococcus* (NCBI accession WP_067098506.1). This Verrumicrobial *ntcA* may occur within this genome as a result of horizontal gene transfer or as an artifact of the MAG binning process. Either way, it indicates that peptide LSHQAIAEAIGSTR should be used with caution as a Cyanobacterial biomarker in environments with abundant Verrumicrobia, and further emphasizes the need to evaluate metagenomic data from the environments analyzed for use of specific peptides as taxonomic biomarkers. However, this is an unlikely scenario as Cyanobacteria tend to be far more abundant in marine environments than Verrumicrobia. Query results from a full sequence NtcA from *Prochlorococcus* MED4 (NCBI GenBank accession CAE18705.1) show that other peptides, such as "LVSFLMVLCR", may be more appropriate if targeting *Prochlorococcus* only (SI Figure S1). By separating sequence types into different categories, METATRYP allows the user to balance the varying levels in confidence of taxonomic attribution, where reference genomes from cultured isolates are the best in taxonomic quality but more incomplete in environmental coverage, and vice versa for environmental sequences.

The addition of 4783 MAGs to METATRYP has provided further insight into the prevalence of shared peptides across taxonomic groups in marine microbial communities. An analysis with METATRYP v 1 using a database of 51 single reference genomes from common pelagic marine microorganisms demonstrated very little overlap in shared peptides

https://dx.doi.org/10.1021/acs.jproteome.0c00385
J. Proteome Res. 2020, 19, 4718−4729

across different taxonomic groups.[1] Expanding this analysis to the 4783 MAGs shows a similar pattern of a very low occurrence of shared tryptic peptides across disparate taxa. Figure 4 shows the individual percentages of shared peptides across a random selection of 25 MAGs. In general, taxa share <1% of tryptic peptides, with a few clusters of more closely related organisms sharing more tryptic peptides—for example, Gammaproteobacterial and Euryarchaeotal clusters highlighted with red outlines—where the taxa share between 1 and 2% or 2 to 14% of tryptic peptides in each cluster, respectively. A similar pattern is shown when the cross-wise comparison of taxa is expanded to 100 MAGs (SI Figures S2 and S3). These results demonstrate that there should be sufficient resolution to discern between taxa using tryptic peptides identified by metaproteomic analyses, especially when coupled to LCA analysis tools like METATRYP to confirm peptide taxonomic origin.

Example of Rapid Deployment: Coronavirus Domain Application

Analysis of the coronavirus-focused METATRYP instance is similar to the observations of marine-focused METATRYP where there is a rather low frequency of shared peptides across disparate taxa with a higher frequency of shared peptides across more closely related taxa (Figure 5; SI Figures S4 and S5). Within the broader group of the 81 taxa associated with severe acute respiratory syndrome (SI Table S4; SI Figures S6 and S7), there is a greater occurrence of shared tryptic peptides. Notably, there is a large cluster of strains associated with the 2003−2004 SARS-CoV-1 outbreak (such as strain Human Coronavirus Frankfurt 1) that cluster together with >90% shared tryptic peptides among the strains. SARS-CoV-2 responsible for the ongoing COVID-19 pandemic is found in a separate cluster among these taxa, clustering most closely with bat-hosted coronaviruses with 84% shared peptides from the SARS-CoV-2 group with the bat RaTG13 virus. The strain SARS Coronavirus Frankfurt 1 (SARS-CoV-1) shares up to 35% of tryptic peptides with the SARS-CoV-2 group. SARS-CoV-2 shares 1% or less of shared tryptic peptides with other taxonomic groups outside the severe acute respiratory syndrome related coronaviruses (Figure 5; SI Figures S6 and S7). This analysis of shared tryptic peptides highlights that SARS-CoV-2 is indeed different, in tryptic peptide space, from other viral pathogens like influenza. Only 1 peptide was identified as shared between SARS-CoV-2 and the major influenza subtypes, sharing peptide sequence "DGQAYVR" from a SARS-CoV-2 spike glycoprotein. However, further inspection of the location of this peptide in influenza and SARS-CoV-2 proteins indicates that this peptide is an artifact of protein purification from sequences used for structural analyses. This sequence is from the C-terminal fibritin of *Escherichia coli* phage T4 used in the heterologous expression of this protein.[43] Therefore, there are no identified tryptic peptides shared between influenza and SARS-CoV-2 as of this analysis date (May 3, 2020). The reference genome for SARS-CoV-2 (NCBI Reference Sequence NC_045512.2) contains 12 protein coding genes with 828 tryptic peptides, thus a SARS-CoV-2 genome shares 0.1% or less tryptic peptides with the influenza pan-proteomes. Thus, differentiating SARS-CoV-2 from other major viral pathogens is tractable using proteomic analyses. Users developing diagnostic assays should take care to independently confirm their informatic analyses, as the METATRYP coronavirus instance is intended as a software example and the current database may not be maintained concurrently with the emergence of new sequencing data.

The coronavirus-focused METATRYP database enables the investigation of shared tryptic peptides across multiple key taxa. A METATRYP database applied in this domain may help identify biomarker peptides that could be used in quantitative proteomics to identify SARS-CoV-2-specific peptides in metaproteomic samples, whether those samples come from human test subjects with complex oral microbiomes or from environmental samples. Using a spike glycoprotein sequence encoded by the S gene from SARS-CoV-2 (NCBI accession YP_009724390.1) as a query, multiple potential SARS-CoV-2 peptides are identified as potential biomarkers for the virus. This protein contains 69 tryptic peptides, with 17 of those peptides being specific to the taxon SARS-CoV-2. Some of the peptides are found across multiple taxa: 52 peptides appear in SARS-CoV-2 and at least one other taxon. Among these, there are peptides which appear to be more conserved across the severe acute respiratory syndrome related viruses: 16 peptides are found in 65 or more Severe acute respiratory related viral taxa (ranging in viruses that infect humans to other hosts like bats, civets, and pigs). The peptide "GIYQTSNFR" may be a potential biomarker for this group of viruses in general, as it is found in all 81 taxa from the severe acute respiratory syndrome related virus group, but not found in other taxa within the database. Given its applications in metaproteomics, METATRYP may be a powerful tool for proteomics applied to SARS-CoV-2 wastewater-based epidemiology used to track community spread of COVID-19 infections,[44,45] as a combination of the coronavirus-specific database with other environmental databases (like marine microbial METATRYP) may provide insight into potential tryptic peptide biomarkers in sewage effluent.

CONCLUSIONS

This manuscript announces the release of version 2 of the METATRYP software package for assessing shared tryptic peptides in complex communities. In addition to the standalone software, we have created web portals for METATRYP v 2 instances that use specialized databases for the marine microbial research community as well as the coronavirus research community. A private API for the marine microbial instance of METATRYP supports LCA analysis in the Ocean Protein Portal[7] and may be further developed into a public API which can be connected to automated pipelines in the future, like GalaxyP.[46] This major release of METATRYP features an upgraded SQL backend that supports faster speeds for database construction and data queries, enables maintenance of separate sequencing data categories within the database, and facilitates LCA analysis of shared peptides across taxa. The main METATRYP web portal database consists of three data categories: "Genomes", "Specialized Assemblies", and "Meta-omic" sequencing assemblies. Users can readily query these databases for the occurrence of specific peptide sequences and visualize the frequency of shared peptides across taxa in the reference "Genomes" and "Specialized Assemblies". The expansion of METATRYP beyond reference "Genomes" allows for more complete coverage of the diversity found in environmental communities; however, these newer sequencing assembly types also carry higher degrees of uncertainty in the taxonomic attributions assigned to the source proteins. The major scientific findings of Saito et al. (2015),[1] that redundancy of tryptic peptides across disparate taxa is rare, is

https://dx.doi.org/10.1021/acs.jproteome.0c00385
J. Proteome Res. 2020, 19, 4718−4729

supported when these new broader sequencing data categories were included in the search database suggesting taxonomic specificity of the majority of tryptic peptides. METATRYP aids in the selection of biomarker peptides for identification of specific taxonomic groups at varying taxonomic levels. We also demonstrated how METATRYP can be applied to proteomics analyses in other scientific domains through the creation of the METATRYP coronavirus web portal. Users can query the occurrence of shared peptides encoded by various coronavirus genomes and other relevant taxa. Using this portal, we showed that the SARS-CoV-2 coronavirus has the most shared tryptic peptides with its closest bat precursor virus, has some shared peptides with SARS-CoV-1, and is very different from the "common flu". METATRYP is a flexible software package to assess taxonomic occurrence of shared peptides applicable to proteomics studies of complex systems valuable for the identification of biomarkers and phyloproteomic analysis of complex communities.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.0c00385.

Table S1: Genomes in the METATRYP web portal and associated taxonomic lineages; Table S2: Metagenome Assembled Genomes (MAGs) in the METATRYP web portal and associated taxonomic lineages; Table S3: Metagenomes included the METATRYP web portal; Table S4: Taxa included in the METATRYP coronavirus web portal and associated taxonomic lineages; Table S5: Subset of genomes included in METATRYP performance benchmarking and the tutorial database and their associated taxonomic lineages; Table S6: Performance metrics reported for METATRYP version 1.1 and version 2.0 database construction and a simple peptide query; Figure S1: Specialized Assembly query results from a full sequence NtcA from *Prochlorococcus* MED4; Figure S2: Heatmap with cluster dendrogram displaying the individual percents of shared tryptic peptides across 100 randomly selected MAGs from the new "Specialized Assembly" data category; Figure S3: Heatmap with cluster dendrogram displaying the individual percents of shared tryptic peptides across 100 randomly selected MAGs from the new "Specialized Assembly" data category with annotated percentages; Figure S4: Heatmap with cluster dendrogram displaying the individual percents of shared tryptic peptides across all taxa currently in the METATRYP coronavirus database; Figure S5: Heatmap with cluster dendrogram displaying the individual percents of shared tryptic peptides across all taxa currently in the METATRYP coronavirus database; Figure S6: Heatmap with cluster dendrogram displaying the individual percents of shared tryptic peptides across all the severe acute respiratory syndrome (SARS) related taxa currently in the METATRYP coronavirus database; Figure S7: Heatmap with cluster dendrogram displaying the individual percents of shared tryptic peptides across all the severe acute respiratory syndrome (SARS) related taxa currently in the METATRYP coronavirus database (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Jaclyn K. Saunders − *Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, United States;* orcid.org/0000-0003-1023-6239; Email: jsaunders@whoi.edu

Mak A. Saito − *Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, United States;* orcid.org/0000-0001-6040-9295; Email: msaito@whoi.edu

Authors

David A. Gaylord − *Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, United States;* orcid.org/0000-0001-7987-6870

Noelle A. Held − *Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, United States;* orcid.org/0000-0003-1073-0851

Nicholas Symmonds − *Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, United States;* orcid.org/0000-0002-9436-0351

Christopher L. Dupont − *J. Craig Venter Institute, San Diego, La Jolla, California 92037, United States*

Adam Shepherd − *Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, United States;* orcid.org/0000-0003-4486-9448

Danie B. Kinkade − *Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, United States;* orcid.org/0000-0002-1134-7347

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.0c00385

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to thank A. Murat Eren, Tom Delmont, Ben Tully, Elaina Graham, and John Heidelberg for graciously providing MAG sequences and additional taxonomic information facilitating incorporation into the METATRYP database. We would also like to thank Alexander J. Devaux for helpful advice in generating the coronavirus database. This work was made possible by grants from the National Science Foundation EarthCube Data Infrastructure Grant NSF-ICER 1639714 and Division of Ocean Science grants NSF-OCE 1657766 and 1924554, the Gordon and Betty Moore Foundation grants 8453 and 3782, and National Institutes of Health (NIH) General Medicine Grant GM135709-01A1. JKS was additionally supported by a NASA Postdoctoral Program Fellowship with the NASA Astrobiology Program, administered by Universities Space Research Association under contract with NASA. METATRYP v 2 is a product of the Ocean Protein Portal (OPP). The OPP team is a collaboration between the Saito laboratory, the Information Services Application group, and the Biological and Chemical Oceanography Data Management Office all at the Woods Hole Oceanographic Institution.

■ REFERENCES

(1) Saito, M. A.; Dorsk, A.; Post, A. F.; McIlvin, M. R.; Rappé, M. S.; DiTullio, G. R.; Moran, D. M. Needles in the blue sea: Sub-species specificity in targeted protein biomarker analyses within the vast oceanic microbial metaproteome. *Proteomics* **2015**, *15* (20), 3521−3531.

(2) Bertrand, E. M.; Moran, D. M.; McIlvin, M. R.; Hoffman, J. M.; Allen, A. E.; Saito, M. A. Methionine synthase interreplacement in diatom cultures and communities: Implications for the persistence of B12 use by eukaryotic phytoplankton. *Limnol. Oceanogr.* **2013**, *58* (4), 1431−1450.
(3) Saito, M. A.; McIlvin, M. R.; Moran, D. M.; Goepfert, T. J.; DiTullio, G. R.; Post, A. F.; Lamborg, C. H. Multiple nutrient stresses at intersecting Pacific Ocean biomes detected by protein biomarkers. *Science* **2014**, *345* (6201), 1173−1177.
(4) Saito, M. A.; McIlvin, M. R.; Moran, D. M.; Santoro, A. E.; Dupont, C. L.; Rafter, P. A.; Saunders, J. K.; Kaul, D.; Lamborg, C. H.; Westley, M.; Valois, F.; Waterbury, J. B. Abundant nitrite-oxidizing metalloenzymes in the mesopelagic zone of the tropical Pacific Ocean. *Nat. Geosci.* **2020**, *13*, 355−362.
(5) Handelsman, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* **2004**, *68* (4), 669−685.
(6) Gurdeep Singh, R.; Tanca, A.; Palomba, A.; Van der Jeugt, F.; Verschaffelt, P.; Uzzau, S.; Martens, L.; Dawyndt, P.; Mesuere, B. Unipept 4.0: Functional Analysis of Metaproteome Data. *J. Proteome Res.* **2019**, *18* (2), 606−615.
(7) Saito, M. A.; Saunders, J. K; Chagnon, M.; Gaylord, D. A.; Shepherd, A.; Held, N. A.; Dupont, C.; Symmonds, N.; York, A.; Charron, M.; Kinkade, D. B. Development of an Ocean Protein Portal for Interactive Discovery and Education. *J. Proteome Res.* **2020**. DOI: 10.1021/acs.jproteome.0c00382
(8) Timmins-Schiffman, E.; May, D. H.; Mikan, M.; Riffle, M.; Frazar, C.; Harvey, H. R.; Noble, W. S.; Nunn, B. L. Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns. *ISME J.* **2017**, *11*, 309.
(9) Saito, M. A.; Bertrand, E. M.; Duffy, M. E.; Gaylord, D. A.; Held, N. A.; Hervey, W. J., IV; Hettich, R. L.; Jagtap, P. D.; Janech, M. G.; Kinkade, D. B.; Leary, D. H.; McIlvin, M. R.; Moore, E. K.; Morris, R. M.; Neely, B. A.; Nunn, B. L.; Saunders, J. K.; Shepherd, A. I.; Symmonds, N. I.; Walsh, D. A. Progress and Challenges in Ocean Metaproteomics and Proposed Best Practices for Data Sharing. *J. Proteome Res.* **2019**, *18* (4), 1461−1476.
(10) Mikan, M. P.; Harvey, H. R.; Timmins-Schiffman, E.; Riffle, M.; May, D. H.; Salter, I.; Noble, W. S.; Nunn, B. L. Metaproteomics reveal that rapid perturbations in organic matter prioritize functional restructuring over taxonomy in western Arctic Ocean microbiomes. *ISME J.* **2020**, *14* (1), 39−52.
(11) May, D. H.; Timmins-Schiffman, E.; Mikan, M. P.; Harvey, H. R.; Borenstein, E.; Nunn, B. L.; Noble, W. S. An Alignment-Free "Metapeptide" Strategy for Metaproteomic Characterization of Microbiome Samples Using Shotgun Metagenomic Sequencing. *J. Proteome Res.* **2016**, *15* (8), 2697−2705.
(12) Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **2012**, *40* (D1), D136−D143.
(13) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422−1423.
(14) McKinney, W. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, Austin, TX, 2010; pp 51−56.
(15) Cantarel, B. L.; Erickson, A. R.; VerBerkmoes, N. C.; Erickson, B. K.; Carey, P. A.; Pan, C.; Shah, M.; Mongodin, E. F.; Jansson, J. K.; Fraser-Liggett, C. M.; Hettich, R. L. Strategies for Metagenomic-Guided Whole-Community Proteomics of Complex Microbial Environments. *PLoS One* **2011**, *6* (11), e27173.
(16) Delmont, T. O.; Quince, C.; Shaiber, A.; Esen, Ö. C.; Lee, S. T. M.; Rappé, M. S.; McLellan, S. L.; Lücker, S.; Eren, A. M. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology* **2018**, *3* (7), 804−813.
(17) Tully, B. J.; Graham, E. D.; Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* **2018**, *5*, 170203−170203.
(18) Alneberg, J.; Bjarnason, B. S.; De Bruijn, I.; Schirmer, M.; Quick, J.; Ijaz, U. Z.; Lahti, L.; Loman, N. J.; Andersson, A. F.; Quince, C. J. N. m. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **2014**, *11* (11), 1144−1146.
(19) Eren, A. M.; Esen, Ö. C.; Quince, C.; Vineis, J. H.; Morrison, H. G.; Sogin, M. L.; Delmont, T. O. J. P. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **2015**, *3*, e1319.
(20) Graham, E. D.; Heidelberg, J. F.; Tully, B. J. J. P. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ* **2017**, *5*, e3035.
(21) Sunagawa, S.; Coelho, L. P.; Chaffron, S.; Kultima, J. R.; Labadie, K.; Salazar, G.; Djahanschiri, B.; Zeller, G.; Mende, D. R.; Alberti, A.; Cornejo-Castillo, F. M.; Costea, P. I.; Cruaud, C.; d'Ovidio, F.; Engelen, S.; Ferrera, I.; Gasol, J. M.; Guidi, L.; Hildebrand, F.; Kokoszka, F.; Lepoivre, C.; Lima-Mendez, G.; Poulain, J.; Poulos, B. T.; Royo-Llonch, M.; Sarmento, H.; Vieira-Silva, S.; Dimier, C.; Picheral, M.; Searson, S.; Kandels-Lewis, S.; Tara Oceans, c.; Bowler, C.; de Vargas, C.; Gorsky, G.; Grimsley, N.; Hingamp, P.; Iudicone, D.; Jaillon, O.; Not, F.; Ogata, H.; Pesant, S.; Speich, S.; Stemmann, L.; Sullivan, M. B.; Weissenbach, J.; Wincker, P.; Karsenti, E.; Raes, J.; Acinas, S. G.; Bork, P.; et al. Ocean plankton. Structure and function of the global ocean microbiome. *Science (Washington, DC, U. S.)* **2015**, *348* (6237), 1261359.
(22) Hoarfrost, A.; Nayfach, S.; Ladau, J.; Yooseph, S.; Arnosti, C.; Dupont, C. L.; Pollard, K. S. Global ecotypes in the ubiquitous marine clade SAR86. *ISME J.* **2020**, *14* (1), 178−188.
(23) Sieracki, M. E.; Poulton, N. J.; Jaillon, O.; Wincker, P.; de Vargas, C.; Rubinat-Ripoll, L.; Stepanauskas, R.; Logares, R.; Massana, R. Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems. *Sci. Rep.* **2019**, *9* (1), 6025.
(24) Markowitz, V. M.; Chen, I. M. A.; Palaniappan, K.; Chu, K.; Szeto, E.; Grechkin, Y.; Ratner, A.; Jacob, B.; Huang, J.; Williams, P.; Huntemann, M.; Anderson, I.; Mavromatis, K.; Ivanova, N. N.; Kyrpides, N. C. IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res.* **2012**, *40* (D1), D115−D122.
(25) Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B. E.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J. B.; Grout, J.; Corlay, S. In *Jupyter Notebooks—A Publishing Format for Reproducible Computational Workflows*; ELPUB, 2016; pp 87−90.
(26) Waskom, M.; Botvinnik, O.; Ostblom, J.; Lukauskas, S.; Hobson, P.; MaozGelbart; Gemperline, D. C.; Augspurger, T.; Halchenko, Y.; Cole, J. B.; Warmenhoven, J.; Ruiter, J. d.; Pye, C.; Hoyer, S.; Verplas, J.; Villalba, S.; Kunter, G.; Quintero, E.; Bachant, P.; Martin, M.; Meyer, K.; Swain, C.; Miles, A.; Brunner, T.; O'Kane, D.; Yarkoni, T.; Williams, M. L.; Evans, C. *mwaskom/seaborn*, v. 0.10.0 (January 2020); DOI: 10.5281/zenodo.3629446.
(27) Wu, F.; Zhao, S.; Yu, B.; Chen, Y.-M.; Wang, W.; Song, Z.-G.; Hu, Y.; Tao, Z.-W.; Tian, J.-H.; Pei, Y.-Y.; Yuan, M.-L.; Zhang, Y.-L.; Dai, F.-H.; Liu, Y.; Wang, Q.-M.; Zheng, J.-J.; Xu, L.; Holmes, E. C.; Zhang, Y.-Z. A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579* (7798), 265−269.
(28) Thiel, V.; Ivanov, K. A.; Putics, Á.; Hertzig, T.; Schelle, B.; Bayer, S.; Weißbrich, B.; Snijder, E. J.; Rabenau, H.; Doerr, H. W.; Gorbalenya, A. E.; Ziebuhr, J. Mechanisms and enzymes involved in SARS coronavirus genome expression. *J. Gen. Virol.* **2003**, *84* (9), 2305−2315.
(29) Bermingham, A.; Chand, M. A.; Brown, C. S.; Aarons, E.; Tong, C.; Langrish, C.; Hoschler, K.; Brown, K.; Galiano, M.; Myers, R.; Pebody, R. G.; Green, H. K.; Boddington, N. L.; Gopal, R.; Price, N.; Newsholme, W.; Drosten, C.; Fouchier, R. A.; Zambon, M. Severe respiratory illness caused by a novel coronavirus, in a patient transferred to the United Kingdom from the Middle East, September 2012. *Eurosurveillance: Eur. Commun. Dis. Bull.* **2012**, *17* (40), 20290.
(30) Killerby, M. E.; Biggs, H. M.; Haynes, A.; Dahl, R. M.; Mustaquim, D.; Gerber, S. I.; Watson, J. T. Human coronavirus

circulation in the United States 2014−2017. *J. Clin. Virol.* **2018**, *101*, 52−56.
(31) van der Hoek, L.; Pyrc, K.; Jebbink, M. F.; Vermeulen-Oost, W.; Berkhout, R. J. M.; Wolthers, K. C.; Wertheim-van Dillen, P. M. E.; Kaandorp, J.; Spaargaren, J.; Berkhout, B. Identification of a new human coronavirus. *Nat. Med.* **2004**, *10* (4), 368−373.
(32) Woo, P. C.; Lau, S. K.; Chu, C. M.; Chan, K. H.; Tsoi, H. W.; Huang, Y.; Wong, B. H.; Poon, R. W.; Cai, J. J.; Luk, W. K.; Poon, L. L.; Wong, S. S.; Guan, Y.; Peiris, J. S.; Yuen, K. Y. Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. *J. Virol.* **2005**, *79* (2), 884−95.
(33) Thiel, V.; Herold, J.; Schelle, B.; Siddell, S. G. Infectious RNA transcribed in vitro from a cDNA copy of the human coronavirus genome cloned in vaccinia virus. *J. Gen. Virol.* **2001**, *82* (6), 1273−1281.
(34) Osada, N.; Kohara, A.; Yamaji, T.; Hirayama, N.; Kasai, F.; Sekizuka, T.; Kuroda, M.; Hanada, K. The genome landscape of the african green monkey kidney-derived Vero cell line. *DNA Res.* **2014**, *21* (6), 673−83.
(35) Jagtap, P.; McGowan, T.; Bandhakavi, S.; Tu, Z. J.; Seymour, S.; Griffin, T. J.; Rudney, J. D. J. P. Deep metaproteomic analysis of human salivary supernatant. *Proteomics* **2012**, *12* (7), 992−1001.
(36) Liévin-Le Moal, V.; Servin, A. L. Anti-infective activities of lactobacillus strains in the human intestinal microbiota: from probiotics to gastrointestinal anti-infectious biotherapeutic agents. *Clin Microbiol Rev.* **2014**, *27* (2), 167−199.
(37) *Identical Protein Groups: Non-redundant Access to Protein Records*; National Center For Biotechnology Information (NCBI), July 26, 2017. https://ncbiinsights.ncbi.nlm.nih.gov/2017/07/26/identical-protein-groups-non-redundant-access-to-protein-records/.
(38) Maglott, D.; Ostell, J.; Pruitt, K. D.; Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research* **2004**, *33* (Database issue), D54−D58.
(39) Sayers, E. *The E-Utilities In-Depth: Parameters, Syntax and More*; National Center for Biotechnology Information (US): Bethesda, MD, 2009.
(40) Ezkurdia, I.; Juan, D.; Rodriguez, J. M.; Frankish, A.; Diekhans, M.; Harrow, J.; Vazquez, J.; Valencia, A.; Tress, M. L. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum. Mol. Genet.* **2014**, *23* (22), 5866−5878.
(41) Zhou, P.; Yang, X.-L.; Wang, X.-G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.-R.; Zhu, Y.; Li, B.; Huang, C.-L.; Chen, H.-D.; Chen, J.; Luo, Y.; Guo, H.; Jiang, R.-D.; Liu, M.-Q.; Chen, Y.; Shen, X.-R.; Wang, X.; Zheng, X.-S.; Zhao, K.; Chen, Q.-J.; Deng, F.; Liu, L.-L.; Yan, B.; Zhan, F.-X.; Wang, Y.-Y.; Xiao, G.-F.; Shi, Z.-L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **2020**, *579* (7798), 270−273.
(42) Herrero, A.; Muro-Pastor, A. M.; Flores, E. Nitrogen control in cyanobacteria. *J. Bacteriol.* **2001**, *183* (2), 411−25.
(43) Wrapp, D.; Wang, N.; Corbett, K. S.; Goldsmith, J. A.; Hsieh, C. L.; Abiona, O.; Graham, B. S.; McLellan, J. S. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **2020**, *367* (6483), 1260−1263.
(44) Mallapaty, S. How sewage could reveal true scale of coronavirus outbreak. *Nature* **2020**, *580* (7802), 176−177.
(45) Mao, K.; Zhang, K.; Du, W.; Ali, W.; Feng, X.; Zhang, H. The potential of wastewater-based epidemiology as surveillance and early warning of infectious disease outbreaks. *Current Opinion in Environmental Science & Health* **2020**, *17*, 1−7.
(46) Jagtap, P. D.; Blakely, A.; Murray, K.; Stewart, S.; Kooren, J.; Johnson, J. E.; Rhodus, N. L.; Rudney, J.; Griffin, T. J. J. P. Metaproteomic analysis using the Galaxy framework. *Proteomics* **2015**, *15* (20), 3553−3565.

https://dx.doi.org/10.1021/acs.jproteome.0c00385
J. Proteome Res. 2020, 19, 4718−4729