UNIVERSITY OF CALIFORNIA SAN DIEGO

What can be learned from Repository-Scale Public Mass Spectrometry Data?

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Computer Science

by

Benjamin Pullman

Committee in charge:

Professor Nuno Bandeira, Chair
Professor Vineet Bafna
Professor Steven Briggs
Professor Pieter Dorrestein
Professor Pavel Pevzner

2022

The Dissertation of Benjamin Pullman is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

This paper is dedicated to my family, my mom, dad, Sarah, Grandma Saunders, Grandma Lulu, Grandpa Pa, Grandpa Eye Eye, who always supported me and pushed me to question everything, to my girlfriend, Samantha, who has both been a source of positivity and a sounding board, to all my teachers who challenged my understanding, and to my music teachers, Ms. Crawford and Moe, who fundamentally made me a better scientist (and probably musician, if I practiced more.)

EPIGRAPH

"In art there are only fast or slow developments. Essentially it is a matter of evolution, not revolution." - Béla Bartók

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Nuno Bandeira for his support as the chair of my committee and all the guidance he has provided in introducing me to the world of computational mass spectrometry and proteomics.

I would also like to acknowledge all the members in the Bandeira Lab who provided support and guidance throughout my PhD as well as those who sat in EBU3B 4252.

Chapter 1, in full, is a reprint of the material as it appears in ProteinExplorer: A Repository-Scale Resource for Exploration of Protein Detection in Public Mass Spectrometry Data Sets in Journal of Proteome Research 2018. Pullman, Benjamin S.; Wertz, Julie; Carver, Jeremy; Bandeira, Nuno, 2018. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in part, is currently being prepared for submission for publication of the material. Pullman, Benjamin S.; Wertz, Julie; Carver, Jeremy; Bandeira, Nuno. The dissertation author was the primary investigator and author of this material.

Chapter 3, in part, is currently being prepared for submission for publication of the material. Pullman, Benjamin S.; Wertz, Julie; Carver, Jeremy; Bandeira, Nuno. The dissertation author was the primary investigator and author of this material.

Chapter 4, in part, is a reprint of the material as it appears in "Real-time modification-tolerant matching of MS/MS spectra at the repository scale". ASMS 2021 - MP159. Pullman, Benjamin S.; Batsoyol, Narangerelt; Swanson, Steven; Wang, Mingxun; Bandeira, Nuno and, in part, is currently being prepared for submission for publication of the material. Pullman, Benjamin S.; Batsoyol, Narangerelt; Swanson, Steven; Wang, Mingxun; Bandeira, Nuno. The dissertation author was the primary investigator and author of this material.

VITA

2013	Bachelor of Arts, Amherst College
2018	Master of Science, University of California San Diego
2022	Doctor of Philosophy, University of California San Diego

ABSTRACT OF THE DISSERTATION

What can be learned from Repository-Scale Public Mass Spectrometry Data?

by

Benjamin Pullman

Doctor of Philosophy in Computer Science

University of California San Diego, 2022

Professor Nuno Bandeira, Chair

High-throughput tandem mass spectrometry has enabled the detection and identification of over 75% of all human proteins predicted to result in translated gene products from an available tens of terabytes of public data in thousands of datasets. This thesis explores what we can learn from this, as well as the challenges that arise when considering proteomics data at a repository scale. First, we will consider validating what is known, through resources to build, curate, and explore both FDR-controlled and user submitted libraries. Second, we present a tool that allows for an automation of application of strict community guidelines criteria to any set of search results, including peak quality and novel FDR controls. Third, we introduce a method to illuminate the extent of what is not yet known using a new clustering approach designed to explicitly model peptide diversity by explicitly modeling spectrum coelutions. Finally, fourth, we developed a method for extremely fast single spectrum searches against spectrum repositories consisting of billions of spectra to both confirm or refute knowledge base IDs as well as discover similar spectra to those consistently unidentified.

Chapter 1

ProteinExplorer: a repository-scale resource for exploration of protein detection in public mass spectrometry datasets

Introduction

Sustained improvements to tandem mass spectrometry (MS/MS) instruments and their application to the analysis of a broad range of protein samples have resulted in the generation of a large volume of mass spectrometry data[1][2]. But while this increased capacity has allowed for the community-wide exploration of the protein content of many types of samples, it has also led to new challenges where the significance of an identification (e.g., peptide or protein) made in the context of a single sample or dataset may not hold when considering aggregate identifications from hundreds of datasets taken to express what is known by the community as a whole. Since the vast majority of the true matches coincide across most datasets (i.e., most human samples share many proteins, which are typically identified mostly by the same peptides) but false matches are much more likely to be unique to each dataset (or at least less consistent across datasets), this leads to a situation where the naïve union of discoveries across many datasets would result in an uncontrolled increase in false discovery rates (FDR) – a problem that affects any question referring to community-scale identifications, as is certainly the case for the HUPO Human Proteome Project's (HPP[3]) quest for reliable detection of translated protein products.

To address this issue, our MassIVE Knowledge Base (MassIVE-KB[4]) spectral library applied strict spectrum, peptide and protein level FDR controls at the aggregate level for all search results included in its construction. As such, its reanalysis of over 31 TB of human HCD data resulted in the largest HCD spectral library to date, with 2.1+ million spectra of $>$1 million unique peptide sequences mapped to $>$16,000 proteins and covering over 50% of all amino acid content in the human proteome. But while MassIVE-KB's identifications offer an FDR-controlled, repository-scale route towards analysis of protein detection in datasets from many sources, it remains cumbersome and time consuming to i) to explore the genomic or functional considerations associated with different identifications and ii) inspect the evidence in support of the detection of missing or dubious proteins whose translated expression has not been confirmed by mass spectrometry data. Our ProteinExplorer application addresses these issues by offering integrated and intuitive access to exon and functional information mapped to peptide and protein identifications, as well as integrating with synthetic peptide[5]/protein[6] expression resources and applying official HPP criteria for detection of novel proteins[3]. In addition, ProteinExplorer provides access to detailed provenance records for every single identification, thereby enabling seamless direct access to the public datasets from which identifications were derived, as well as to the standardized search jobs that were used to generate the original identifications from the public mass spectrometry data – an aspect that is often overlooked when aggregating results from multiple datasets (i.e., access to the full set of search parameters and original search results). In particular, the hundreds of public datasets in the current release of ProteinExplorer reported here already include identifications from systematic reanalysis of multiple datasets that were incorporated into previous HPP releases (PXD000529/MSV000080255[7], PXD000533/ MSV000080254[7], PXD000561/MSV000079514[8], PXD000612/MSV000080701[9], PXD000865/MSV000079526[10], PXD003947/MSV000080826[11]), thereby allowing for open access to inspect whether reanalysis confirms the original reports after considering repository-scale FDR corrections.

To further support the validation of peptide and protein identifications beyond what is supported in existing resources, ProteinExplorer facilitates the comparison of experimental spectra identified from public datasets to spectra of synthetic peptides (from the ProteomeTools[5] project) or proteins (from the BioPlex[6] project) by including these in separate spectral libraries (altogether covering over 19,000 proteins) and by providing a simple, one-click access to the generation of interactive spectrum images for matching peptides in multiple libraries. Finally, since there are multiple ways in which identifications may not be currently eligible to fully claim detection of novel proteins (e.g., protein identifications with only one peptide or absence of corroborating spectra of synthetic peptides), ProteinExplorer further incorporates a user library entitled PrEdict (Protein Existence dictionary) where users can submit spectra in partial support of the detection of novel proteins, and thus iteratively progress towards collaborative detection of proteins across datasets. Illustrating how this feature supports community inspection of evidence of detection and supports convergence towards a common consensus, the quality and interpretation of these identifications has also been rated and sometimes commented in records associated with each entry in the user library.

Methods

ProteinExplorer is built on the MassIVE Knowledge Base[4] (MassIVE-KB), a repository-scale spectral library resulting from the reanalysis of 658 million MS/MS HCD spectra from 27,404 LC/MS runs in 227 datasets. In brief, spectra were searched with MS-GF+[12] database search against the UniProt human reference proteome with isoforms as well as contaminants, such as porcine trypsin. Variable modifications included in the searches were oxidation, N-term acetylation, N-term Carbamylation, Pyro-glu, and deamidation were considered as variable modifications and carbamidomethylation on Cysteine as a fixed modification. For the purpose of the analysis reported here, only matches identified to canonical proteins, and not matches to contaminants, isoforms, or TrEMBL were retained. Spectrum identifications were considered to be ambiguous and removed from consideration if there were two or more peptides matching the spectrum with scores passing FDR thresholds. FDR was applied at the spectrum level (0.1% library-level FDR), length-specific peptide level (1% local FDR), and 1% protein-level FDR for all proteins matched by at least one unique peptide[13][14], which corresponds to 0.013% protein-level FDR for proteins matched by 2+ peptides[15] (i.e., expect two false positive protein identifications). In addition to the MassIVE-KB spectral library constructed from public datasets of natural sequences, spectral libraries of synthetic peptide spectra were also constructed using the same process and used to assess the correlation of fragmentation patterns with MassIVE-KB spectra.

Protein metadata and functional information was integrated into ProteinExplorer from both PhosphoSitePlus[16] and Uniprot[17], with the current release based on downloads on March 26, 2018. Protein Existence (PE) classifications were downloaded from the neXtProt[18] ftp (the 2018-01-17 release) and incorporated into ProteinExplorer as well (rare discrepancies between UniProtKB and neXtProt due to lags in synchronization between database versions were conservatively assigned a PE of 0 to remove those matches from consideration). In brief, protein existence tiers are defined as PE1 if there is evidence at protein level, PE2 if evidence at transcript level, PE3 if inferred from homology, PE4 if predicted and PE5 if uncertain; proteins are also referred to as "missing" if classified as PE2, PE3 or PE4 and "dubious" if classified as PE5[18]. To further assess the potential biological significance of peptides, all peptides were mapped onto genomic exons included in reference human transcripts. Using a previously described approach[16][17], we mapped all peptides in MassIVE-KB to exons in ENSEMBL 89[19][20] and annotated whether the peptide maps uniquely to an exon, covers a splice junction, or is mapped to an exon at all (some peptides are not mapped due to differences between UniProt and ENSEMBL sequences). The algorithms used to determine whether proteins were matched by two or more HPP-compliant peptides are described in Supplementary Materials.

ProteinExplorer is developed as a community hub for examining the human proteome and is built as a Java application with a JavaScript front-end and a RESTful API built with Tomcat to fit in the ecosystem of the MassIVE repository and the ProteoSAFe parallel workflow engine; ProteinExplorer can be accessed at https://massive.ucsd.edu/ProteoSAFe/protein_explorer_splash.jsp. The underlying database is built in MySQL and contains tables for libraries, proteins, peptides, representatives (spectral library PSMs selected to best characterize a modified peptide at a given precursor charge out of the set of all PSMs which pass FDR), provenance spectra (the PSMs that pass FDR for each precursor), and comments. In particular, PSMs are stored in a way that is independent of libraries, so when updates are made to a library, or even representatives, it is unlikely that most PSMs will need to be updated. An overview of the ProteinExplorer functionality and data sources is provided in Figure 1.1 and the full schema for the database is provided in Figure 1.5.

The JavaScript user interface provides two views for exploring the proteome. The first is a proteome-wide view with information about all proteins, which we call the proteome page, and the second is a protein-centric view that contains amino-acid level coverage information from libraries as well as metadata from community resources, we which title the protein page. The proteome page consists of two panels. The first is a series of filters for protein accession identifiers, as well as other common fields such as the Uniprot protein description, protein existence (PE) information from neXtProt, and options to consider only specific public datasets. Under these filter boxes is a table that displays all proteins that satisfy the search criteria specified in the filter boxes. The table also provides a one-click option to separate protein information per dataset, including dataset specific expression and peptide-uniqueness; the table is also downloadable for offline processing (see supplementary materials for examples of how the proteome view can be utilized).

The protein page provides a detailed view of each protein, including sequence coverage, representatives, and provenance, as well as filters for library expression and functional information overlays. This page is specific per protein and can also be filtered to show only combinations of libraries (e.g. only the natural peptides). The coverage map superimposes sequence-level expression for the protein with functional information from UniProt and PhosphoSitePlus, and is interactive allowing for users to hover over amino acids for more information about metadata as well as click on individual amino acids to filter the view to just peptides covering that amino acid. As the interface displays ample information in a relatively small space, distinctions are made between different libraries by using different colors, which are fully customizable to be more accessible to all users, using a color picker[21]. When two libraries overlap, we blend the colors by simply averaging the RGB values for each color, allowing for an easy distinction between red, blue, and their mix (purple). The final two subsections of the protein page provide interactive and downloadable tables for the representative peptides for the library and for all the supporting PSM provenance information for the protein. The representatives table includes information about representative length, number of proteins that each peptide maps onto when considering single amino acid variants, number of exons matched by each peptide, and the coordinates of the peptide on the protein. The provenance table contains all information necessary to track the peptide back to its original experiment as well as search task, including the filename, scan, charge, search algorithm, and link to the search. These tables are all filterable, and any filter made to any table is applied to the entire view, facilitating discovery by eliminating redundant clicks.

Algorithm for finding HUPO proteins

Beyond simply listing the PE classification for every human protein, ProteinExplorer was also designed to facilitate the assessment of HUPO Extraordinary Detection Claims[22] by pre-calculating the required criteria for a novel protein to be considered detected: it must contain two non-nested peptides of length 9 or greater, where each is uniquely mapped to the undigested sequence of proteins from the same gene even when considering all human proteins (and their isoforms) in the reference proteome, within a single amino acid variant (SAAV) of the peptide sequence. To calculate the number of non-nested peptides, we maintain all the peptides that only map uniquely to a given protein, and then sort all of these peptides for the protein by their end coordinate, End[i], in increasing order and for cases where the end coordinate matches, their start coordinate, Start[i], in increasing order as well. We then formulate a dynamic programming algorithm with the following recurrence, MaxPeps[i] represents the number of maximum number of non-overlapping peptides when including peptide i as part of the solution and MaxScoreAtPep[i] represents the maximum number of non-overlapping peptides when using peptide candidates up to and including peptide i, regardless of whether peptide i is included in the solution or not. The maximum number of non-overlapping peptides will be at the final position in MaxScoreAtPep.

Synthetics

In addition to the MassIVE-KB spectral library constructed from public datasets of non-synthetic sequences, spectral libraries of synthetic peptide spectra were also constructed using a similar process and used to assess the correlation of fragmentation patterns with MassIVE-KB spectra. The first batch of synthetic peptide spectra was from the ProteomeTools project[5] (PXD004732/MSV000080544) and the second batch of synthetic peptide spectra was from the BioPlex project[6] (MSV000080679), considering only peptides from synthetic sequences in runs where those were used as bait proteins for affinity purification mass spectrometry (AP-MS); in addition, since there was evidence of carry-over between consecutive BioPlex mass spectrometry runs, peptide identifications matching the bait sequence from the previous run were also considered to be from synthetic sequences and were not included in the set of identifications considered for detection of novel proteins.

To calculate similarity between a MassIVE-KB representative spectrum and a spectrum from a synthetic peptide, we consider the maximum cosine between all MassIVE-KB PSMs identified to the peptide sequence and all spectra of synthetic peptides assigned to the same peptide sequence; before calculation of cosines, all spectra were preprocessed to remove precursor ion peaks (and their neutral losses), as well as filtered to retain only peaks with intensity rank of at least 10 in a window of +/-50 Da around the peak mass. In the one case where the observed cosine was lower than what we would expect for spectral library matches at 1% false discovery rate (at which most usually accepted high-resolution matches feature cosines of 0.6 or higher), we conservatively manually inspected the peptide-spectrum match to confirm the agreement of fragmentation between the synthetic and non-synthetic peptide spectra. See Figure 1.9 for the spectra that are matched.

But since FDR thresholds, stringent as they may be, are still insufficient to guarantee the correctness of every single identification (i.e., the error estimates do not indicate which identifications are incorrect), ProteinExplorer also includes spectral libraries of synthetic peptides to further establish trust in the identifications derived from MassIVE-KB's large-scale analysis of public proteomics data. As described before[4] and in the Methods section, ProteinExplorer spectral libraries of synthetic peptides were derived from the ProteomeTools collection of synthetic peptides5 ("ProteomeTools Synthetics" library under "Library selection") and from the BioPlex collection of AP-MS experiments[6], where synthetic genes were originally used to analyze the protein interactions of the translated gene products ("BioPlex Synthetics" library under "Library selection"). As illustrated in Figure 1.3a in the main text, since peptide sequences matched to spectra in each library are shown in different colors on the protein coverage display, it becomes straightforward to visually identify protein regions that are covered by peptides from only one library (e.g., only synthetic peptides or only non-synthetic peptides). Conversely, visual rendering of peptides from different libraries mapping to overlapping protein regions is achieved by superimposing the colors selected for the corresponding libraries. By default, since the MassIVE-KB no synthetics library (in-vivo expression) is shown in red, and ProteomeTools and BioPlex are shown in blue, all regions matched by both in vivo and synthetic peptides are thus shown in purple.

Sorting by either "Start AA" or by "Sequence" reveals pairs of PSMs assigned to the same sequence but with representative PSMs from either "MassIVE-KB (no synthetics)" (in vivo expression) or from synthetic peptides in either "ProteomeTools Synthetics" or in "BioPlex Synthetics" (see Figure 1.3b in the main text). Finally, as illustrated in Figure 1.3c in the main text, clicking on the spectrum icon on the leftmost column shows the MS/MS spectra supporting each peptide identification, with fragment peaks annotated according to the assigned sequence and with an interactive display allowing for comparison to theoretical masses, as well as other features such as zooming in/out, consulting experimental peak masses and considering alternative explanations for the assigned peptide sequence (by changing the sequence atop the spectrum image and clicking “Update Peptide).

Results

The high diversity of peptide sequences and expression patterns observed across many datasets in repository-scale searches open up the possibility of investigating protein biology in more detail by exploring patterns of protein expression across datasets (see Supplementary Materials) as well as comprehensive coverage of genomic exons and functional sites. As shown in Figure 1.2, multiple features are mapped onto a sequence coverage view of the protein sequence and it is possible to highlight different aspects of the coverage by selecting between different types of rendering of the protein sequence, with “Flat coverage” highlighting all amino acids covered by at least one PSM, “Spectrum Coverage” highlighting amino acids on a color scale based on how many PSMs covered it and “Peptide Coverage” and “Variants Coverage” highlighting amino acids based on how many unique peptides or modified peptide variants cover each site (respectively). This coverage view is then extended to facilitate the analysis of exon matches and functional sites.

Peptide sequences are mapped to genomic coordinates[23][24] on ENSEMBL 89[19][20] to i) determine whether the mapping is unique (and thus indicative of exon presence/absence) and ii) whether the peptide is fully contained within an exon or it’s a junction peptide spanning one or more exons. As such, filtering by “Unique Exon Match” or by “Exon Junction Match” in the “Peptide Representatives” table (i.e., by requiring a minimum value of one on either of these columns) would select only for peptides matching the corresponding category and update the sequence coverage to highlight only the locations covered by the sequences of the selected peptides. Sorting by “Start AA” or “End AA” in the “Peptide Representatives” table further facilitates inspection of overlapping sequences and modification variants covering the same protein regions.

Functional knowledge of modified sites was integrated from UniProt[17] and Phospho-

SitePlus[16] are shown on the protein coverage view as underlines, with additional details shows by hovering over annotated sites. These be further explored to examine if any of the peptides contain modified variants of a site or whether modification or expression patterns vary across a range of datasets. Clicking on the amino acid location in the protein coverage panel filters the page for only sequences covering the site of interest. Upon filtering, the views renormalize the expression levels to examine specific sites rather than global expression.

Beyond the sequence level understanding in the coverage view, the “Peptide representatives” table then provides additional information on these peptides including the diversity of peptide sequences and modification variants covering the site and the MS/MS spectrum, as well as their corresponding expression patterns in terms of both number of PSM identifications (“Provenance Spectra” column) and number of datasets in which each peptide variant was detected (“Dataset Occurrences” column), thereby facilitating the estimation of which peptides are most representative of the protein (e.g., identified in the highest number of datasets), rather than just peptides with large number of PSMs (which might also be explained by dataset-specific protocols strongly selecting for some protein regions over others).

We illustrate this functionality by considering a disease-associated site annotated by PhosphoSitePlus on protein P06733 (Figure 1.2) and examining the overlapping variants at position 2[21]. As all views are filtered by clicking on this site, we see that there are 4058 PSMs in 114 variants that cover this site. Further filtering for acetylated peptide sequences by entering "+42.011" in the filter box for the "Sequence" field in the "Peptide Representatives" table reveals that, even though the original searches did not consider Lysine acetylation as a possible modification, there were still 9 peptide variants covering this site and identified as modified with N-terminal acetylation. Out of these, 3 variants account for over 90% of all observed PSMs and are the most commonly observed across multiple datasets. Interestingly, selecting for Asparagine deamidation (by filtering for "N+0.984") also reveals separate deamidation on each of the three Asparagines on the most abundant peptide sequence covering the site (DATNVGDEGGFAPNILENK), each of which is clearly supported by site-localizing peaks, as can be readily seen by clicking on the spectrum icon on the leftmost column of the peptide representatives table.

ProteinExplorer also facilitates the design of targeted experiments. The standard course of action in a targeted experiment is to first find a few unique precursors to the protein and then find transitions, (i.e. precursor and ion pairs) that are frequently observed in a reproducible manner. In the protein view it is possible to click on a site to see all covering peptides which can then be further filtered to emphasize particular characteristics, e.g. selecting for a particular dataset that might be illustrative of experimental conditions. Once a peptide is selected, one can then rank PSMs and compare them side by side to pick transitions (see Supplementary Materials).

Detection of PE2-4 and PE5 missing proteins

The inspection of evidence provided in support of the detection of PE2-4 and PE5 proteins (missing and uncertain/dubious proteins) is a core need of the Human Proteome Project's (HPP) goal of confirming the in vivo translation of proteins predicted from the human genome. Establishing the detection of these proteins requires that the supporting identifications meet a well-defined set of criteria for extraordinary detection claims[22], including verification of the identifications using spectra of synthetic peptides, all of which are addressed in ProteinExplorer as follows.

First, the requirements for proper estimation of false discovery rates at the spectrum, peptide and protein levels are all guaranteed by the spectral library construction processes used to assemble the MassIVE-KB spectral library from over 190 million PSMs identified from over 30 TB of human higher-energy collisional dissociation (HCD) data4. Second, the additional requirements for peptide length of at least nine amino acids, unique protein matches even with one single amino acid variation (SAAV), and detection of at least two non-contained peptides were addressed in ProteinExplorer and are reflected in the protein page in the "# Matched Proteins w/0-1 SAAV Mismatch". As such, filtering on this column for peptides matching exactly one protein selects for all candidates potentially supporting the detection of missing proteins.

The final step in the confirmation of peptide identifications (matching fragmentation to spectra of synthetic peptides) is implemented in ProteinExplorer by separately listing representative PSMs with matching sequences but from different libraries (the “Library” field in the “Peptide Representatives” table). Further, clicking on the “View Only Overlaps” option under the “Coverage Type” options (as shown in Figure 1.3) refines the protein coverage to render only peptides matched by two or more libraries, thereby facilitating the visual selection of sequences. Also in this view, it is possible to track every single PSM back to the scan number, raw file and public dataset where the spectral data came from (data provenance), as well as trace every PSM back to the search workflow, parameters, job and results from which the PSM was extracted (analytical provenance). Both of these elements are critical to the evaluation of detection of missing proteins and likely should be added as formal requirements for all HPP submissions.

Using all of these ProteinExplorer features to analyze identifications in the current release of MassIVE-KB, we were able to detect 296 PE2-4 and 55 PE5 proteins with at least one HPP-compliant peptide, with 107 missing (PE2, PE3 and PE4) and 23 dubious (PE5) proteins of these identified with two or more peptides meeting all HPP criteria for peptide identifications (see Figure 1.4a). We include PE5 proteins as we have found ample evidence for the detection of many but continue to report them separately from the missing proteins. Out of these PE2-4 proteins, 60 (3 for PE5) had two or more peptides with matching spectra from synthetic peptides and 21 (4 for PE5) proteins had only one of the two peptides for which spectra of synthetic peptides were available (see Table S2 which is split into two sheets for both missing and dubious proteins and is further sortable by number of matching synthetics on the “HPP Non-overlapping Matching Synthetics” column as well as containing chromosome location and a link to the ProteinExplorer page where all MS/MS spectra for the peptides can be examined). The distribution of cosines between MassIVE-KB representative spectra (identified from public datasets) and corresponding spectra of synthetic peptides is shown in Figure 1.4b; spectra and rationale for approval are provided in supplementary materials as well as can be browsed online for all matches with a cosine less than 0.6 for further examination (see Supplementary Materials and Figure 1.9). While spectra of synthetic peptides were not available to validate all 107 PE2-4 and 23 PE5 protein identifications with 2+ HPP-compliant peptides, we note that out of 60*2+21=141 PE2-4 and 3*2+4=10 PE5 cases where matching sequences were present in both MassIVE-KB and synthetic peptide collections, all spectrum matches confirmed the identifications – thereby strongly suggesting that the vast majority of proteins currently detected with 2+ HPP-compliant peptides will also be confirmed as soon as spectra of corresponding synthetic peptides become available.

In addition to facilitating the confirmation of HPP criteria for detection of missing proteins, ProteinExplorer also helps ascertain whether proteins are likely to ever generate enough peptides meeting the current criteria. For example, protein H3BUK9 is currently identified in MassIVE-KB by 26 distinct peptide variants (and further matched an additional 34 shared peptides) and is matched by over 200 variants in ProteomeTools and Bioplex synthetics combined, altogether accumulating 15 peptide sequences (of length 9 or longer) uniquely matching to this protein – yet not a single one of these peptides is a unique match after considering a single amino acid variation (see Figure 1.2).

Updating libraries with user-added spectra

Since many proteins in the human proteome have not yet been detected or have been matched only to a degree that does not meet HPP guidelines, ProteinExplorer also supports contributions to a user library entitled PrEdict (Protein Existence dictionary) constituting a community resource of exploratory ‘peptide hints‘ that may eventually accumulate to the point of fully supporting the detection of previously unobserved proteins. Submissions to the PrEdict library can come from complete dataset submissions or from reanalyses of public datasets, even if these have not yet been aggregated into repository-scale spectral libraries (or otherwise guaranteed to meet repository-scale FDR thresholds, as is the case with MassIVE-KB spectral library construction workflows)). That said, ProteinExplorer retains the full provenance of every

PSM by linking it back to the dataset and specific context from where it was contributed (e.g., a dataset reanalysis), thereby providing a route for direct inspection of the quality of search producing the PSM.

To demonstrate the potential utility of the ProteinExplorer PrEdict library, we added CID spectra from the CPTAC colorectal dataset[25] (MSV000079852) in support of an additional 25 missing proteins, each matched by 2 non-overlapping peptides matching uniquely to the protein sequence even when considering SAAVs (see Table S4). These PSMs were originally either submitted with the dataset or identified by subsequent CCMS reanalysis using MS-GF+ database search. As an example of the utility of manual revision of tentative spectra, we added a peptide ARHSEAEATRAR that uniquely maps (including SAAV matches) to the protein A6NNA2, a protein with 8 synthetic peptides (125 PSMs) but only a limited amount of observations in natural data (1 peptide with 1 PSM). While the spectrum initially appears to have a fair amount of unexplained ion current, it turns out that the highest intensity peaks can be explained as doubly-charged ions for neutral losses from the unfragmented precursor. As such, the spectrum was marked as a 4 star match and a comment was entered (and is directly attached to the spectrum for immediate access by anyone inspecting it further) describing the reasons for the rating. All spectra in all ProteinExplorer libraries can be rated and annotated in the same way, thereby empowering the community to review all matches submitted in support of the detection of missing proteins.

Entry point and global searches

Access to large-scale protein identifications usually starts with looking for proteins matching criteria considered in the ProteinExplorer main query page, from where it is possible to search for specific proteins with an accession or gene name, as well as by UniProt protein description (see Figure 1.7). In addition, it is also possible to select for proteins by their neXtProt PE level (e.g., all PE2 proteins predicted from known transcripts) or by dataset (e.g., all proteins identified in the MSV000079514/PXD000561 reanalysis of the draft human proteome dataset).

Searching by UniProt protein description facilitates selection of proteins from a common subset; for example, querying for the most commonly expressed HLA protein can be achieved by searching for “HLA-” in the protein description box to get a list of the 101 proteins that are HLA related and have evidence in MassIVE-KB (see Figure 1.7a). The list can then be sorted and further filtered by several additional fields shown in the protein list table, such as sorting by decreasing number of peptides uniquely mapped to each protein (ranging from 86 to 0 in this case), as well as by PE level, revealing that almost all PE genes are classified as PE1 with hundreds to thousands of peptide spectrum matches (PSMs) revealing that the most observed HLA protein in these results was P10316. The only exception with PE>1 was P01893 / HLA-H which is classified as PE5, but clicking to “view only non-synthetic matches” promptly reveals that even though the protein is identified with 3 unique non-synthetic peptide matches, none of them meets the HPP criteria for detection of novel proteins. All results on the protein list table can also be downloaded for further offline processing by clicking the “Export Filtered Results” button immediately above the table header.

Exploring protein expression across datasets is also enabled by simply clicking the option to “view dataset protein pairs”, which can combine with filtering for a specific protein to see its varying expression across datasets. For example, filtering for the HLA protein with most unique peptides (P01903, with 83 unique non-synthetic peptides) shows that it is detected in 55 datasets but with broad variation in the number of unique peptides identified per dataset (e.g., after sorting by decreasing “unique peptides per dataset”), each of which has substantially fewer unique peptides than the total number observed across all datasets. Upon closer inspection of the datasets with the most unique peptides for this protein, it emerges that while some of these datasets correspond to similar protocols (e.g., following the typical trypsin digestion protocols), others focus on less-common protocols that are likely to reveal a different peptide space, such as neoepitopes (MSV000080517) or endogenous peptides (MSV000080859), thereby suggesting that the selection of most representative peptides for a protein may depend on the tissue or type of sample and protocols being used for data acquisition. Further filtering the list to get only matches to the draft proteome datasets (dataset accessions can be filtered using comma-separated lists of datasets such as “MSV000079526,MSV000079514”) shows that the number of unique peptides and exon matches observed in these reference datasets is more comparable, as is expected from their similar tissue sources and experimental protocols. In contrast, the comparison with MSV000080517 also reveals that the number of unique peptides does not always increase in proportion to the numbers of PSMs, with the focus on neoepitopes capturing nearly twice as many unique peptides while identifying less than half as many PSMs as the draft proteomes.

Designing targeted experiments

The design of targeted experiments to detect proteins and protein features (e.g., isoforms or modification variants) across conditions or experiments[25] can be informed by the repository-scale diversity of expression patterns observed across all datasets. Once a protein of interest is selected, ProteinExplorer facilitates the selection of peptides to consider for targeted experiments in a variety of ways.

If the user wants to develop a targeted experiment to see the expression of a protein, the standard course of action is to first find a few unique representatives to that protein and then find transitions, or precursor and ion pairs, that are frequently observed in a reproducible manner. Simply setting the upper bound of the “#Matched Proteins” column to 1 initiates this process by selecting only peptides that match uniquely to the protein; this also ensures that all subsequent analysis is done for only these unique peptides (since the filters are applied to the entire page).

The list of representative peptides can then be sorted by number of PSMs using the “Provenance Spectra” column to get a sense for the relative detectability of the multiple peptides from the same protein. Alternatively, the representatives can also be sorted by decreasing number of datasets where they were observed using the “Dataset Occurrences” column, which can be used as an indication of consistency of observation under varying samples and experimental conditions. Furthermore, ProteinExplorer can also highlight regions of the coverage map where there is a high amount of diversity of peptide modification variants by setting the coverage to “Variant

Level", thereby highlighting protein regions more prone to (likely artefactual) modifications which could complicate higher accuracy of protein quantification. Also, since the colors for the coverage map are averaged in areas where the sequence is matched by peptides from multiple libraries, the user could hone their search to include only variants that occur in multiple libraries, and potentially have spectra of matched synthetic peptides.

Once a peptide is selected, peaks for the transitions must be decided upon. By clicking to view the spectrum of the peptide representative, the user can find and select the most intense peaks simply by visual inspection. However, when selecting transitions, it can also be useful to look beyond the representative to consider a larger collection of PSMs for the same peptide to assess consistency of peptide fragmentation, as well as possible inconsistent interferences. The additional PSMs can be found using the "PSM Provenance" information right below the representatives, where its spectrum images can show which peaks are consistently observed across many PSMs and datasets. Further, the peaks can be compared to those in spectra of synthetic peptides to confirm whether the ions and relative intensities are maintained between synthetic and non-synthetic peptides.

Beyond developing experiments for targeting proteins, ProteinExplorer also facilitates the application of the same methods for targeting functional sites. As shown above, to view all representatives covering a functional site, the user simply has to click that functional site. From here, a similar process can be taken, first looking at detectability and consistency for all peptides covering the site of interest, and then deciding on a representative. If modifications are involved, it could also be informative to first sort by PSM counts and then find the most common version of the representative where the site of interest may or may not be modified, and then compare the fragmentation patterns of the modified and unmodified peptide to select the most informative peaks for analysis of the modified site. Similarly, experiments can also be developed to look at genomic architecture such as exon-unique or exon-junction peptides. Currently, we provide a supplementary table (Table S3) that shows a mapping from peptides to proteins/exons which can be used to decide the peptides to target (e.g., by selecting peptides that are “Exon Unique” or “Exon Junction”). From these peptides, the user can then filter the “Peptide Representatives” table on “Sequence” to find the location on the protein coverage and, by filtering the representatives to cover the sequence of interest, allow for an understanding of the diversity of sequence and modification variants in the peptide region as well as an understanding of the PSMs and datasets where the exon or exon junction occurs.

Process to update libraries with user-added spectra

Peptide identifications can be added to the user library by clicking the “Add to Library” link in any mzTab results view for any MassIVE dataset or reanalysis at http://massive.ucsd.edu/. The reanalysis could be previously submitted or a newly run by the user. Clicking on the “Add to Library” link redirects to an input form wherein the user can confirm the details of the selected peptide, including resolution of modifications masses to Unimod accessions. Submitted peptides are then mapped to the proteome to find all coordinates in all proteins matching the peptide sequence and a SAAV-tolerant search is conducted over the entire proteome to determine whether the peptide can be marked as an ‘HPP peptide’ and potentially used as supporting evidence for detection of a novel protein. After completing the submission, the user is then redirected to the protein page showing where the peptide was mapped and whether it overlaps with peptides from other libraries (including potential matches to synthetic peptides); see Figure 1.8 for an overview of the process.

Since submission of spectra to the user library is expected to be tentative in that the protein identifications are uncertain (and peptide identifications have not passed repository-scale FDR), it is important to allow for these PSMs to be rated and commented on by the community on their potential to eventually stand in support of the detection of novel proteins. As such, users can rate PSMs in any ProteinExplorer library on a scale of 1 to 4 stars, with 1 star indicating incorrect match with a companion comment indicating alternative explanation, 2 stars indicating likely incorrect match but without a companion alternative explanation (e.g., based on fragmentation properties), 3 stars indicating partially correct identifications (e.g., incorrect site localization for post-translational modification) and 4 stars indicating agreement with reported identification.

Discussion

The identification of over 1.4 million modified human peptide variants from systematic reanalysis of public mass spectrometry data has created new opportunities for understanding the human proteome, but also brings new challenges for the meaningful and efficient exploration of such a large number of identifications from a volume of data so large that would not be accessible to the vast majority of proteomics labs. ProteinExplorer thus expands beyond other resources in i) its ability to enable interactive exploration of multiple libraries in dynamically updated views displaying expression and functional metadata and ii) by supporting the submission and curation of spectrum identifications for missing/dubious proteins, thereby empowering the community to add spectra to the PrEdict user library, as well as curate (i.e., rate and comment on) other spectra which were also submitted as hints.

As a key example of the utility of ProteinExplorer to assist with the exploration of the human proteome, we have also described how its features enable the inspection of evidence submitted in support of the detection of novel proteins, and show how these select identifications for dozens of proteins that are fully compliant with HPP guidelines, as well as detect many more proteins whose peptide identifications match HPP guidelines but for which there are currently no spectra of the same synthetic peptide sequences. Since it is thus expected that it will be common for evidence in support of the detection of novel proteins to be accumulated over time as new data becomes available (including data for spectra of synthetic peptides), ProteinExplorer supports this iterative and collaborative process by allowing for the submission of PSMs to the PrEdict user library that is shared by the whole community. Finally, progressing towards community-wide consensus of which identifications to accept requires full transparency in complete provenance records (from data, tools and search procedures, all the way to identifications) but should also be supported by an interactive platform where (dis)agreements can be recorded and used to eventually converge on a community-curated collection of reliably identified spectra, especially for cases of high biological relevance (e.g., binding regions in monoclonal antibodies) or for evidence supposed to support the detection of novel proteins.

Acknowledgements

Chapter 1, in full, is a reprint of the material as it appears in ProteinExplorer: A Repository-Scale Resource for Exploration of Protein Detection in Public Mass Spectrometry Data Sets in Journal of Proteome Research 2018. Pullman, Benjamin S.; Wertz, Julie; Carver, Jeremy; Bandeira, Nuno, 2018. The dissertation author was the primary investigator and author of this paper.

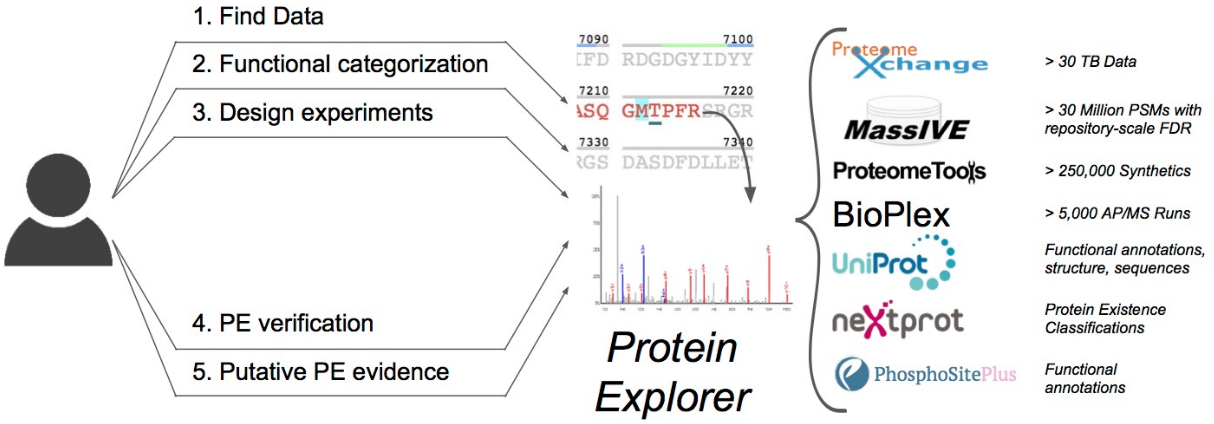
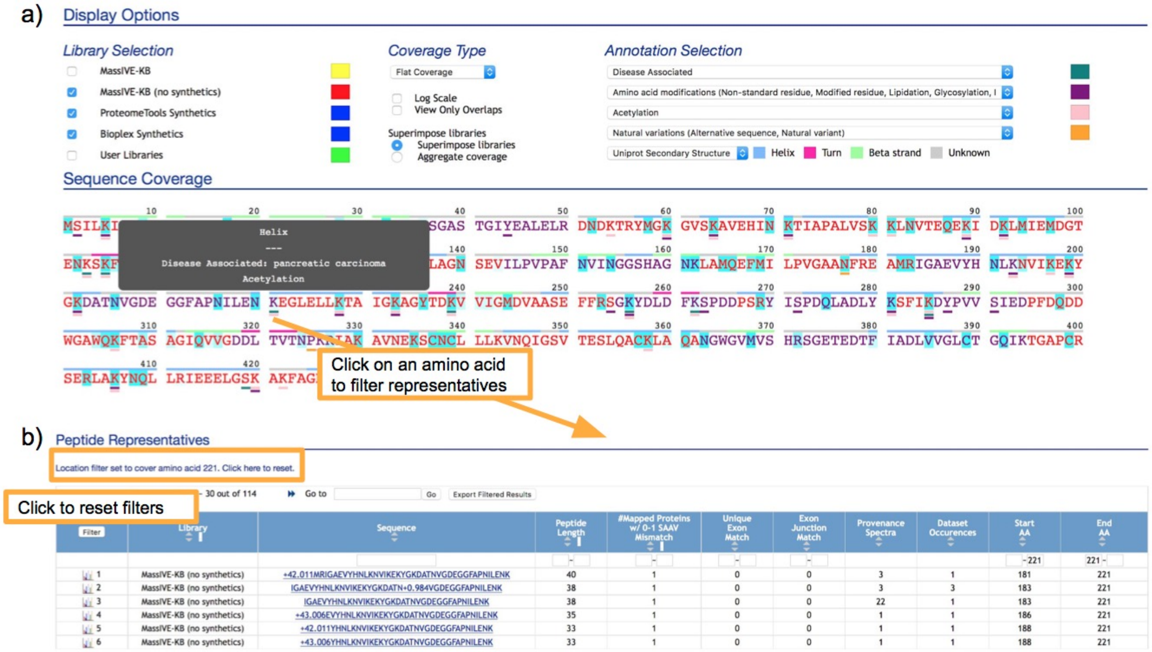


Figure 1.1. ProteinExplorer was designed to facilitate the productive exploration of repository-scale identification of tens of millions of peptide spectrum matches (PSMs) for over 1 million distinct peptide sequences identified from 30+ TB of public mass spectrometry data.



a) Display Options

Library Selection

□ MassIVE-KB 🟨
☑ MassIVE-KB (no synthetics) 🟥
☑ ProteomeTools Synthetics 🟦
☑ Bioplex Synthetics 🟦
□ User Libraries 🟩

Coverage Type

Flat Coverage

□ Log Scale
□ View Only Overlaps

Superimpose libraries
● Superimpose libraries
○ Aggregate coverage

Annotation Selection

Disease Associated
Amino acid modifications (Non-standard residue, Modified residue, Lipidation, Glycosylation, I
Acetylation
Natural variations (Alternative sequence, Natural variant)
Uniprot Secondary Structure 🟦 Helix 🟥 Turn 🟩 Beta strand ⬜ Unknown

Sequence Coverage

```
         10         20         30         40         50         60         70         80         90        100
MSILKI                                  SGAS TGIYEALELR DNDKTRYMGK GVSKAVEHIN KTIAPALVSK KLNVTEQEKI DKLMIEMDGT
                                         140        150        160        170        180        190        200
ENKSKF                                  LAGN SEVILPVPAF NVINGGSHAG NKLAMQEFMI LPVGAANFRE AMRIGAEVYH NLKNVIKEKY
                                         240        250        260        270        280        290        300
GKDATNVGDE GGFAPNILEN KEGLELLKTA IGKAGYTDKV VIGMDVAASE FFRSGKYDLD FKSPDDPSRY ISPDQLADLY KSFIKDYPVV SIEDPFDQDD
       310        320        330        340        350        360        370        380        390        400
WGAWQKFTAS AGIQVVGDDL TVTNPK  AK AVNEKSCNCL LLKVNQIGSV TESLQACKLA QANGWGVMVS HRSGETEDTF IADLVVGLCT GQIKTGAPCR
       410        420
SERLAKYNQL LRIEEELGSK AKFAG
```

Helix

Disease Associated: pancreatic carcinoma
Acetylation
Click on an amino acid
to filter representatives

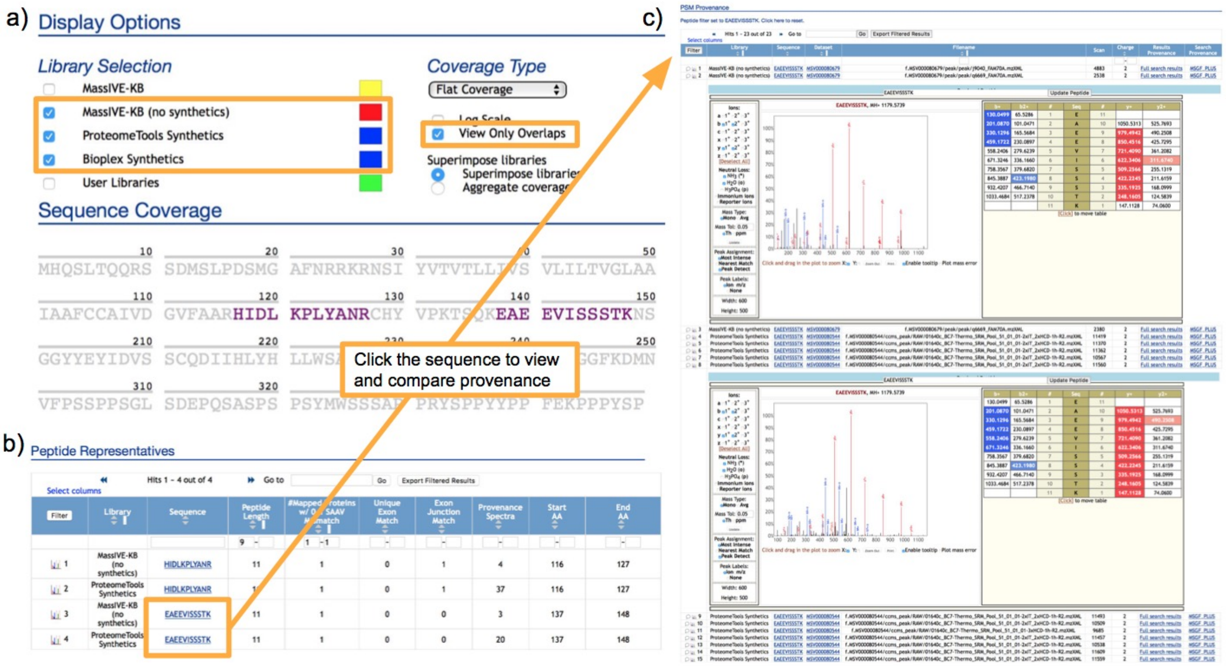
b) Peptide Representatives

Location filter set to cover amino acid 221. Click here to reset.

- 30 out of 114 ⏩ Go to [ ] Go | Export Filtered Results

Filter	Library	Sequence	Peptide Length	#Mapped Proteins w/ 0-1 SAAV Mismatch	Unique Exon Match	Exon Junction Match	Provenance Spectra	Dataset Occurences	Start AA	End AA
			-	-	-	-			-221	221-
1	MassIVE-KB (no synthetics)	+42.011MRIGAEVYHNLKNVIKEKYGKDATNVGDEGGFAPNILENK	40	1	0	0	3	1	181	221
2	MassIVE-KB (no synthetics)	IGAEVYHNLKNVIKEKYGKDATN+0.984VGDEGGFAPNILENK	38	1	0	0	3	3	183	221
3	MassIVE-KB (no synthetics)	IGAEVYHNLKNVIKEKYGKDATNVGDEGGFAPNILENK	38	1	0	0	22	1	183	221
4	MassIVE-KB (no synthetics)	+43.006EVYHNLKNVIKEKYGKDATNVGDEGGFAPNILENK	35	1	0	0	1	1	186	221
5	MassIVE-KB (no synthetics)	+42.011YHNLKNVIKEKYGKDATNVGDEGGFAPNILENK	33	1	0	0	1	1	188	221
6	MassIVE-KB (no synthetics)	+43.006YHNLKNVIKEKYGKDATNVGDEGGFAPNILENK	33	1	0	0	1	1	188	221

Figure 1.2. (a) For each protein, the sequence coverage display provides an easy view to explore the spectrum, peptide and modified variants coverage with superimposed metadata from UniProt and PhosphoSitePlus. In this example, we highlight UniProt amino acid modifications and disease associated sites from PhosphoSitePlus as dashes below their associated site, with secondary structure also shown above the protein sequence. (b) Modified peptide variants covering a site that is disease-associated (from PhosphoSitePlus) and also a known modification (from UniProt), as well as covered by modified peptide variants in MassIVE-KB.



a) Display Options

Library Selection

□ MassIVE-KB 🟨
☑ MassIVE-KB (no synthetics) 🟥
☑ ProteomeTools Synthetics 🟦
☑ Bioplex Synthetics 🟦
□ User Libraries 🟩

Coverage Type

Flat Coverage

□ Log Scale
☑ View Only Overlaps

Superimpose libraries
● Superimpose libraries
○ Aggregate coverage

Sequence Coverage

```
         10         20         30         40         50
MHQSLTQQRS SDMSLPDSMG AFNRRKRNSI YVTVTLLIVS VLILTVGLAA
        110        120        130        140        150
IAAFCCAIVD GVFAARHIDL KPLYANRCHY VPKTSQKEAE EVISSSTKNS
        210        220                              250
GGYYEYIDVS SCQDIIHLYH LLWSA                   GGFKDMN
        310        320
VFPSSPPSGL SDEPQSASPS PSYMWSSSAP PRYSPPYYPP FEKPPPYSP
```

Click the sequence to view and compare provenance

b) Peptide Representatives

Hits 1 - 4 out of 4 Go to [Go] [Export Filtered Results]

Select columns

[Filter]	Library	Sequence	Peptide Length	#Mapped Proteins w/ 0-1 SAAV Mismatch	Unique Exon Match	Exon Junction Match	Provenance Spectra	Start AA	End AA
			9 -	1 - 1	-	-	-	-	-
1	MassIVE-KB (no synthetics)	HIDLKPLYANR	11	1	0	1	4	116	127
2	ProteomeTools Synthetics	HIDLKPLYANR	11	1	0	1	37	116	127
3	MassIVE-KB (no synthetics)	EAEEVISSSTK	11	1	0	0	3	137	148
4	ProteomeTools Synthetics	EAEEVISSSTK	11	1	0	0	20	137	148

c) PSM Provenance

Peptide filter set to EAEEVISSSTK. Click here to reset.

Hits 1 - 23 out of 23 Go to [Go] [Export Filtered Results]

Select columns

[Filter]	Library	Sequence	Dataset	Filename	Scan	Charge	Results Provenance	Search Provenance
1	MassIVE-KB (no synthetics)	EAEEVISSSTK	MSV000080679	f.MSV000080679/peak/peak/j9040_FAM70A.mzXML	4883	2	Full search results	MSGF_PLUS
2	MassIVE-KB (no synthetics)	EAEEVISSSTK	MSV000080679	f.MSV000080679/peak/peak/q6669_FAM70A.mzXML	2538	2	Full search results	MSGF_PLUS

EAEEVISSSTK [Update Peptide]

EAEEVISSSTK, MH+ 1179.5739

b+	b2+	#	Seq	#	y+	y2+
130.0499	65.5286	1	E	11		
201.0870	101.0471	2	A	10	1050.5313	525.7693
330.1296	165.5684	3	E	9	979.4942	490.2508
459.1722	230.0897	4	E	8	850.4516	425.7295
558.2406	279.6239	5	V	7	721.4090	361.2082
671.3246	336.1660	6	I	6	622.3406	311.6740
758.3567	379.6820	7	S	5	509.2566	255.1319
845.3887	423.1980	8	S	4	422.2245	211.6159
932.4207	466.7140	9	S	3	335.1925	168.0999
1033.4684	517.2378	10	T	2	248.1605	124.5839
		11	K	1	147.1128	74.0600

Figure 1.3. (a) ProteinExplorer facilitates comparison of spectra from different libraries (e.g., for comparison of spectra from synthetic and natural sources) by providing an option to automatically select only peptides with the same sequence in the two libraries; e.g., selecting ProteomeTools and natural MassIVE-KB to allow for inspection of full compliance with HPP criteria. (b) List of selected peptides with entries in multiple libraries, also displaying whether the peptides are sufficiently long and do not match more than one protein even when allowing for single amino acid variations (SAAVs). (c) Interactive Lorikeet panels render annotated PSMs for assessment of matches to theoretical ions masses, as well as for inspection of correlated fragmentation between spectra from synthetic and natural sources.

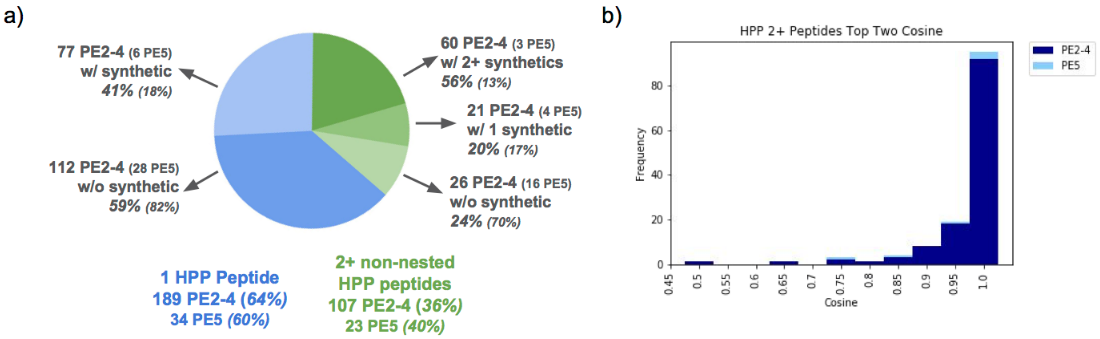


Figure 1.4. (a) MassIVE-KB identifications with repository-scale FDR detect 365 protein annotated by neXtProt at the level of protein existence 2 or higher (PE2+), out of which we find that 63 PE2+ proteins were identified with 2+ non-nested peptides whose spectra matched to spectra of synthetic peptides. (b) Cosine distribution for synthetic/natural matches for all peptides supporting the detection of PE2+ proteins; all cases with cosine below 0.6 were manually inspected and are shown and discussed in supplementary materials.

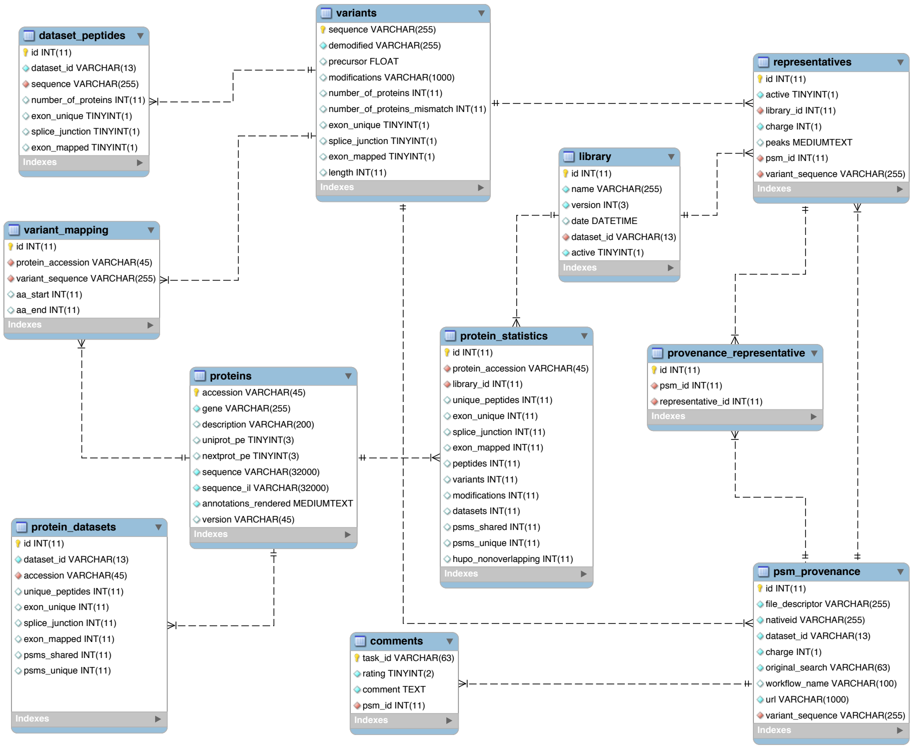
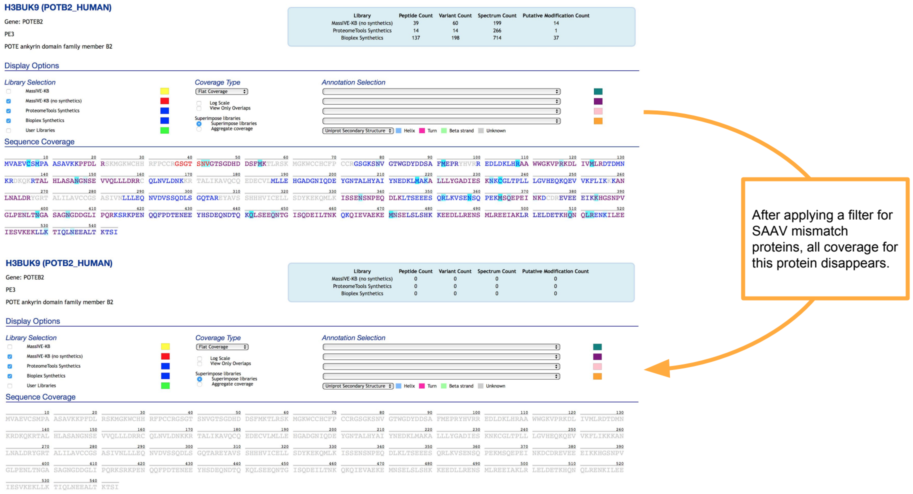


Figure 1.5. Database Schema. Complete schema of the MySQL database of ProteinExplorer.



H3BUK9 (POTB2_HUMAN)

Gene: POTEB2

PE3

POTE ankyrin domain family member B2

Library	Peptide Count	Variant Count	Spectrum Count	Putative Modification Count
MassIVE-KB (no synthetics)	39	60	199	14
ProteomeTools Synthetics	14	14	266	1
Bioplex Synthetics	137	198	714	37

Display Options

Library Selection

- □ MassIVE-KB 🟨
- ☑ MassIVE-KB (no synthetics) 🟥
- ☑ ProteomeTools Synthetics 🟦
- ☑ Bioplex Synthetics 🟦
- □ User Libraries 🟩

Coverage Type

Flat Coverage

- □ Log Scale
- □ View Only Overlaps

Superimpose libraries

- ● Superimpose libraries
- ○ Aggregate coverage

Annotation Selection

Uniprot Secondary Structure ▨ Helix ▨ Turn ▨ Beta strand ▨ Unknown

Sequence Coverage

```
        10         20         30         40         50         60         70         80         90        100        110        120        130
MVAEVCSMPA ASAVKKPFDL RSKMGKWCHH RFPCCRGSGT SNVGTSGDHD DSFMKTLRSK MGKWCCHCFP CCRGSGKSNV GTWGDYDDSA FMEPRYHVRR EDLDKLHRAA WWGKVPRKDL IVMLRDTDMN
       140        150        160        170        180        190        200        210        220        230        240        250        260
KRDKQKRTAL HLASANGNSE VVQLLLDRRC QLNVLDNKKR TALIKAVQCQ EDECVLMLLE HGADGNIQDE YGNTALHYAI YNEDKLMAKA LLLYGADIES KNKCGLTPLL LGVHEQKQEV VKFLIKKKAN
       270        280        290        300        310        320        330        340        350        360        370        380        390
LNALDRYGRT ALILAVCCGS ASIVNLLLEQ NVDVSSQDLS GQTAREYAVS SHHHVICELL SDYKEKQMLK ISSENSNPEQ DLKLTSEEES QRLKVSENSQ PEKMSQEPEI NKDCDREVEE EIKKHGSNPV
       400        410        420        430        440        450        460        470        480        490        500        510        520
GLPENLTNGA SAGNGDDGLI PQRKSRKPEN QQFPDTENEE YHSDEQNDTQ KQLSEEQNTG ISQDEILTNK QKQIEVAEKE MNSELSLSHK KEEDLLRENS MLREEIAKLR LELDETKHQN QLRENKILEE
       530        540
IESVKEKLLK TIQLNEEALT KTSI
```

H3BUK9 (POTB2_HUMAN)

Gene: POTEB2

PE3

POTE ankyrin domain family member B2

Library	Peptide Count	Variant Count	Spectrum Count	Putative Modification Count
MassIVE-KB (no synthetics)	0	0	0	0
ProteomeTools Synthetics	0	0	0	0
Bioplex Synthetics	0	0	0	0

Display Options

Library Selection

- □ MassIVE-KB 🟨
- ☑ MassIVE-KB (no synthetics) 🟥
- ☑ ProteomeTools Synthetics 🟦
- ☑ Bioplex Synthetics 🟦
- □ User Libraries 🟩

Coverage Type

Flat Coverage

- □ Log Scale
- □ View Only Overlaps

Superimpose libraries

- ● Superimpose libraries
- ○ Aggregate coverage

Annotation Selection

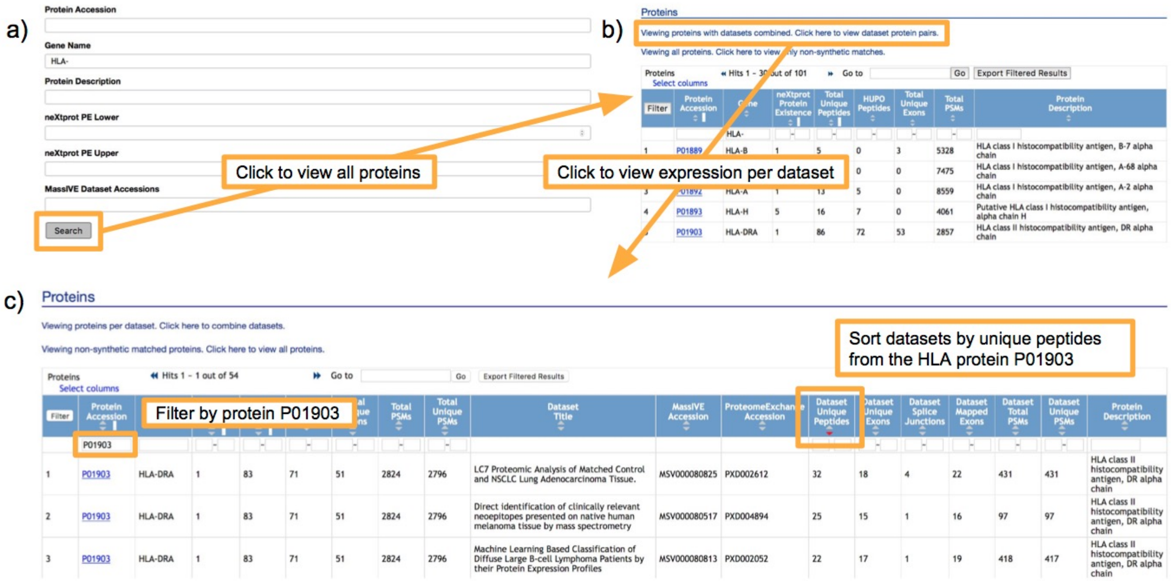
Uniprot Secondary Structure ▨ Helix ▨ Turn ▨ Beta strand ▨ Unknown

Sequence Coverage

```
        10         20         30         40         50         60         70         80         90        100        110        120        130
MVAEVCSMPA ASAVKKPFDL RSKMGKWCHH RFPCCRGSGT SNVGTSGDHD DSFMKTLRSK MGKWCCHCFP CCRGSGKSNV GTWGDYDDSA FMEPRYHVRR EDLDKLHRAA WWGKVPRKDL IVMLRDTDMN
       140        150        160        170        180        190        200        210        220        230        240        250        260
KRDKQKRTAL HLASANGNSE VVQLLLDRRC QLNVLDNKKR TALIKAVQCQ EDECVLMLLE HGADGNIQDE YGNTALHYAI YNEDKLMAKA LLLYGADIES KNKCGLTPLL LGVHEQKQEV VKFLIKKKAN
       270        280        290        300        310        320        330        340        350        360        370        380        390
LNALDRYGRT ALILAVCCGS ASIVNLLLEQ NVDVSSQDLS GQTAREYAVS SHHHVICELL SDYKEKQMLK ISSENSNPEQ DLKLTSEEES QRLKVSENSQ PEKMSQEPEI NKDCDREVEE EIKKHGSNPV
       400        410        420        430        440        450        460        470        480        490        500        510        520
GLPENLTNGA SAGNGDDGLI PQRKSRKPEN QQFPDTENEE YHSDEQNDTQ KQLSEEQNTG ISQDEILTNK QKQIEVAEKE MNSELSLSHK KEEDLLRENS MLREEIAKLR LELDETKHQN QLRENKILEE
       530        540
IESVKEKLLK TIQLNEEALT KTSI
```

After applying a filter for SAAV mismatch proteins, all coverage for this protein disappears.

Figure 1.6. H3BUK9 is a protein that has many natural peptide matches in ProteinExplorer, but they are all shared with other proteins.



Protein Accession

Gene Name: HLA-

Protein Description

neXtprot PE Lower

neXtprot PE Upper

MassIVE Dataset Accessions

Search

Click to view all proteins

Proteins

Viewing proteins with datasets combined. Click here to view dataset protein pairs.

Viewing all proteins. Click here to view only non-synthetic matches.

Proteins « Hits 1 - 30 out of 101 » Go to | Go | Export Filtered Results

Select columns

Click to view expression per dataset

Filter	Protein Accession	Gene	neXtprot Protein Existence	Total Unique Peptides	HUPO Peptides	Total Unique Exons	Total PSMs	Protein Description
		HLA-						
1	P01889	HLA-B	1	5	0	3	5328	HLA class I histocompatibility antigen, B-7 alpha chain
					0	0	7475	HLA class I histocompatibility antigen, A-68 alpha chain
3	P01892	HLA-A	1	13	5	0	8559	HLA class I histocompatibility antigen, A-2 alpha chain
4	P01893	HLA-H	5	16	7	0	4061	Putative HLA class I histocompatibility antigen, alpha chain H
	P01903	HLA-DRA	1	86	72	53	2857	HLA class II histocompatibility antigen, DR alpha chain

Proteins

Viewing proteins per dataset. Click here to combine datasets.

Viewing non-synthetic matched proteins. Click here to view all proteins.

Sort datasets by unique peptides from the HLA protein P01903

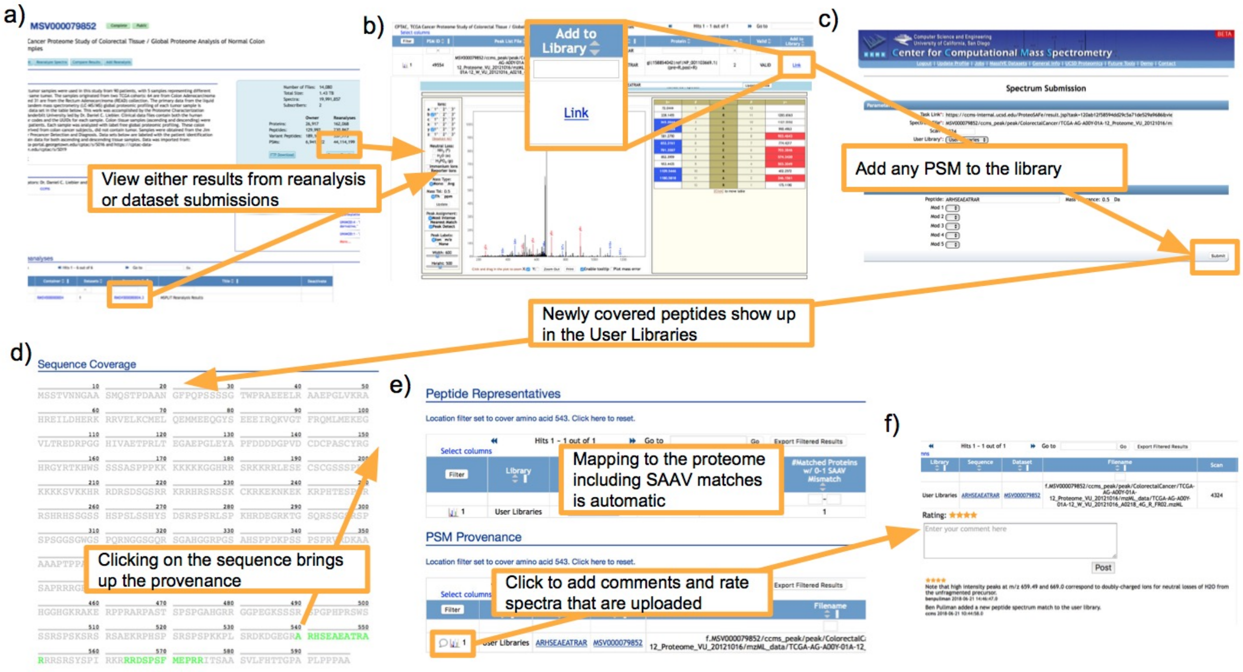
Proteins ◀◀ Hits 1 - 1 out of 54 ▶▶ Go to | Go | Export Filtered Results

Select columns

Filter by protein P01903

Filter	Protein Accession					al ue ns	Total PSMs	Total Unique PSMs	Dataset Title	MassIVE Accession	ProteomeExchange Accession	Dataset Unique Peptides	Dataset Unique Exons	Dataset Splice Junctions	Dataset Mapped Exons	Dataset Total PSMs	Dataset Unique PSMs	Protein Description
	P01903																	
1	P01903	HLA-DRA	1	83	71	51	2824	2796	LC7 Proteomic Analysis of Matched Control and NSCLC Lung Adenocarcinoma Tissue.	MSV000080825	PXD002612	32	18	4	22	431	431	HLA class II histocompatibility antigen, DR alpha chain
2	P01903	HLA-DRA	1	83	71	51	2824	2796	Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry	MSV000080517	PXD004894	25	15	1	16	97	97	HLA class II histocompatibility antigen, DR alpha chain
3	P01903	HLA-DRA	1	83	71	51	2824	2796	Machine Learning Based Classification of Diffuse Large B-cell Lymphoma Patients by their Protein Expression Profiles	MSV000080813	PXD002052	22	17	1	19	418	417	HLA class II histocompatibility antigen, DR alpha chain

Figure 1.7. (a) The ProteinExplorer main page allows users to search for proteins by accession, gene, neXtprot PE, and UniProt description. (b) The protein list panel shows all proteins meeting the filtration criteria, either at the level of the whole repository or per dataset. (c) The dataset option allows users to see and filter by dataset accessions to find dataset-specific protein information, including unique peptide and exon counts.



View either results from reanalysis or dataset submissions

Add any PSM to the library

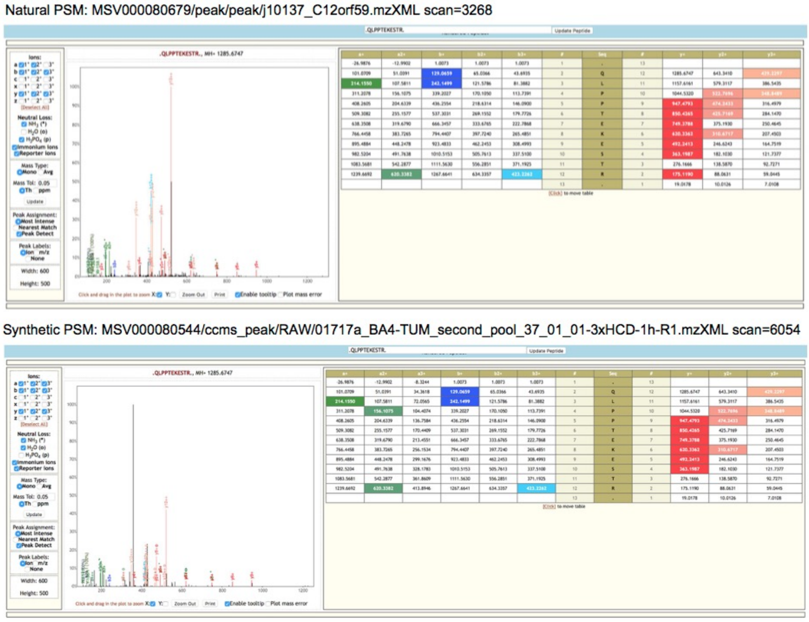
Newly covered peptides show up in the User Libraries

Clicking on the sequence brings up the provenance

Mapping to the proteome including SAAV matches is automatic

Click to add comments and rate spectra that are uploaded

Figure 1.8. ProteinExplorer allows for user submission to shared community libraries. While user libraries are not FDR-controlled in the same way as MassIVE-KB libraries, the flexibility to share spectra of any PSMs facilitates the exchange of information and could help converge towards collaborative accumulation of evidence supporting the detection of novel proteins. (a) All public datasets are potential sources for submission of spectra to user libraries. (b) Dataset results or reanalyses allow for inspection of PSMs and submission to the user libraries. (c) The spectrum submission form curates the PSM for resolution of modifications (if needed) before final submission to the library. (d) Submitted PSMs are mapped onto the proteome and shown in both the coverage map and in the (e) representative and provenance sections. (f) Spectra in the user library (and in all other libraries) can be rated and commented on by the community to assess and promote consensus around the proposed identifications.



Natural PSM: MSV000080679/peak/peak/j10137_C12orf59.mzXML scan=3268

.QLPPTEKESTR. Update Peptide

a+	a2+	b+	b2+	b3+	#	Seq	#	y+	y2+	y3+
-26.9876	-12.9902	1.0073	1.0073	1.0073	1	.	13			
101.0709	51.0391	129.0659	65.0366	43.6935	2	Q	12	1285.6747	643.3410	429.2297
214.1550	107.5811	242.1499	121.5786	81.3882	3	L	11	1157.6161	579.3117	386.5435
311.2078	156.1075	339.2027	170.1050	113.7391	4	P	10	1044.5320	522.7696	348.8489
408.2605	204.6339	436.2554	218.6314	146.0900	5	P	9	947.4793	474.2433	316.4979
509.3082	255.1577	537.3031	269.1552	179.7726	6	T	8	850.4265	425.7169	284.1470
638.3508	319.6790	666.3457	333.6765	222.7868	7	E	7	749.3788	375.1930	250.4645
766.4458	383.7265	794.4407	397.7240	265.4851	8	K	6	630.3362	315.6717	207.4503
895.4884	448.2478	923.4833	462.2453	308.4993	9	E	5	492.2413	246.6243	164.7519
982.5204	491.7638	1010.5153	505.7613	337.5100	10	S	4	363.1987	182.1030	121.7377
1083.5681	542.2877	1111.5630	556.2851	371.1925	11	T	3	276.1666	138.5870	92.7271
1239.6692	620.3382	1267.6641	634.3357	423.2262	12	R	2	175.1190	88.0631	59.0445
					13	.	1	19.0178	10.0126	7.0108

[Click] to move table

Synthetic PSM: MSV000080544/ccms_peak/RAW/01717a_BA4-TUM_second_pool_37_01_01-3xHCD-1h-R1.mzXML scan=6054

.QLPPTEKESTR. Update Peptide

a+	a2+	a3+	b+	b2+	b3+	#	Seq	#	y+	y2+	y3+
-26.9876	-12.9902	-8.3244	1.0073	1.0073	1.0073	1	.	13			
101.0709	51.0391	34.3618	129.0659	65.0366	43.6935	2	Q	12	1285.6747	643.3410	429.2297
214.1550	107.5811	72.0565	242.1499	121.5786	81.3882	3	L	11	1157.6161	579.3117	386.5435
311.2078	156.1075	104.4074	339.2027	170.1050	113.7391	4	P	10	1044.5320	522.7696	348.8489
408.2605	204.6339	136.7584	436.2554	218.6314	146.0900	5	P	9	947.4793	474.2433	316.4979
509.3082	255.1577	170.4409	537.3031	269.1552	179.7726	6	T	8	850.4265	425.7169	284.1470
638.3508	319.6790	213.4551	666.3457	333.6765	222.7868	7	E	7	749.3788	375.1930	250.4645
766.4458	383.7265	256.1534	794.4407	397.7240	265.4851	8	K	6	630.3362	315.6717	207.4503
895.4884	448.2478	299.1676	923.4833	462.2453	308.4993	9	E	5	492.2413	246.6243	164.7519
982.5204	491.7638	328.1783	1010.5153	505.7613	337.5100	10	S	4	363.1987	182.1030	121.7377
1083.5681	542.2877	361.8609	1111.5630	556.2851	371.1925	11	T	3	276.1666	138.5870	92.7271
1239.6692	620.3382	413.8946	1267.6641	634.3357	423.2262	12	R	2	175.1190	88.0631	59.0445
						13	.	1	19.0178	10.0126	7.0108

[Click] to move table

Figure 1.9. These are two spectra from the same variant of a PE2 protein Q4KMG9 that has 2+ matching HPP peptides, but the cosine match for this example is lower than expected at 0.461. However, this score can be explained due to a high intensity contaminant peak, though otherwise the y-series y4+ through y9+ matches and many other fragmentations peaks correlate.

Chapter 2

HPP-Inspector: automated community-scale validation of novel protein discoveries

Introduction

Proteomics discoveries are increasingly supported by large datasets, including draft proteome datasets such as PXD010154[26] and PXD016999[27], which consider dozens of different tissues to develop a comprehensive view of the proteome. Rare discoveries are also enabled by smaller-scale datasets focused on less-explored tissues, samples, or experimental protocols too. These data have been fundamental to the Human Proteome Project (HPP) run by the Human Proteome Organization (HUPO) is the largest community-scale proteomics data science project to date, aiding in the goal of determining the existence and biological context of protein products for every human gene. Since many datasets explore the same (or very similar) biological samples, it is often the case that true discoveries are replicated across datasets whereas false discoveries resulting from experimental variations are likely to be less consistent, thus resulting in increased false discovery rates (FDRs). Recognizing this challenge, the HPP introduced and frequently revised a set of guidelines that should be followed to establish the reliability of protein identifications, especially in cases aiming to establish protein existence by providing first-time evidence of detection of in vivo protein expression. The manual application of the HPP guidelines has served the community well by reinforcing the robustness of the evidence provided in support of protein existence but its broader impact has been limited by the lack of tools for automating their application to a broader scope and scale of search results. HPP-Inspector was developed to automate the application of HPP guidelines to any search results from any public dataset to a sharable, high-level summary of potential novel proteins, while allowing complete provenance back to the dataset and the original search to validate the claims.

A key component of the reanalysis pipeline is approaches for assessing spectrum quality or robustness of identifications. PeptideProphet[28] and Percolator[29] use spectrum attributes to rescore PSMs, but the HPP guidelines require "high signal to noise” identifications directly and may be diluted among other criteria in these approaches, whereas HPP-Inspector measures this directly. Quality control approaches such as those recorded in qcML[30] are more focused on determining the viability of an entire run, and are not focused on specifically understanding single spectrum signal-to-noise. Determining HPP criteria at the sequence level for the guidelines is also important, and the NeXtProt Uniqueness checker[31], does this by reports if a given input sequence is unique to a given protein to satisfy HPP criteria, but does not consider spectrum quality. Finally, while ProteinExplorer[32] also implements the HPP-matching tools, it is designed to work only on highly curated data and requires time consuming steps to add new datasets.

In difference from other tools, HPP-Inspector also implements both spectrum quality rescoring as well as sequence-based criteria such considering SAAVs and determining the maximal-scoring set of non-redundant peptides of length no less than 9 amino acids. HPP-Inspector also implements comparison to synthetics, which are generally understood to be the gold standard in validating identification. In addition to implementing criteria to filter for high-signal to noise identifications and to filter out possible single amino acid variations (SAAVs), HPP-Inspector also implements automated comparison of identifications to spectra of synthetic peptides and allows for the comparison of different popular approaches for calculation of protein FDR. Finally, HPP-Inspector also allows for the comparison of new search results to identifications already provided in publicly available knowledge bases (such as the MassIVE-KB

Knowledge Base[4]) to help assess the potential new contributions from new datasets and search results.

Methods

Data inputs

HPP-Inspector is a workflow in the ProteoSAFe/MassIVE environment, accessible at massive.ucsd.edu; the open-source code is also available on GitHub at https://github.com/CCMS-UCSD/hpp_inspector (v1.12 at time of submission.) The first required input to the workflow is a set of peptide identifications, preferably in the open mzTab[33] PSI format (e.g., as included in many ProteomeXchange complete dataset submissions) or from MassIVE-KB libraries (described in "Running MassIVE-KB libraries through HPP-Inspector"). mzTab conversion workflows are also available at MassIVE for identification files in either mzIdentML[34] or TSV formats containing at least columns for MS run, spectrum identifier, peptide identification, and PSM-score (see supplementary information for direct links to all workflows). Protein identifications are not required since HPP-Inspector remaps all peptide sequences. The column to use for Peptide Spectrum Match (PSM) scores can be selected on the HPP-Inspector input form, as well as whether higher (e.g., number of matched ions) or lower (e.g., p-value) scores correspond to better-scoring PSMs. For search results using public data available at MassIVE, spectrum peaks will be automatically loaded from the MassIVE repository; otherwise, spectrum files can be uploaded to a MassIVE user account and used as inputs to the workflow. The second required input to the workflow is a UniProt FASTA file containing the protein sequences and with protein headers containing the substring “GN=<gene(s)>” listing the genes that each protein sequence is transcribed from. It is recommended to use UniProt Reference Proteomes whenever possible; results presented here were obtained using the UniProt human reference human proteome, release 2020-06[35] All sequences from the input are extracted and mapped to the input proteome FASTA, allowing for a single amino acid mismatch (SAAV) for each peptide, both expected and unexpected[36] using a previously described algorithm[32][4].

Spectral libraries containing reference spectra of synthetic peptides can also be provided as input to assess the correlation of their MS/MS fragmentation with that of the input PSMs. The file format supported for spectral libraries is a simple extension of the standard MGF format (see supplementary information for details), as is currently implemented for releases of the MassIVE-KB[4] spectral library. Multiple synthetic inputs can be considered for the same precursor, and the highest similarity synthetic will be considered for the match.

Spectrum processing

Not all spectra that pass 1% PSM level FDR are high enough quality to be evidence for a protein and HPP-Inspector automatically filters out low quality PSMs as well as matches variants to synthetics, ensuring that spectra selected for further analysis are of sufficient quality, complete details in Supplementary Information.

Global False Discovery Rate (FDR)

While the inputs are expected at 1% PSM level FDR, to ensure that the proteins we are considering meet HPP criteria, they must also be controlled at protein level FDR as well. In HPP-Inspector there are four ways to consider FDR. The first, traditional FDR, uses the summed intensity of all variants, where a variant is defined as having the same sequence and summed modification masses differing by $<=$ 3 Daltons, that pass quality filters and uniquely map to a canonical or contaminant protein. Contaminant proteins are included in this calculation as these represent true positives of proteins existing in the matrix, even if not necessarily human. FDR is then calculated for each protein, P, by considering the number of decoys divided by number of targets for all proteins with score greater than or equal to the score of P. Picked FDR is also considered, to avoid over-representation of decoys as is common in large datasets[15], and that is calculated using the same input variants (see "FDR").

However, the HPP criteria also state that each protein must have 2+ non-nested peptides that are uniquely mapping to a protein, even when considering SAAVs. Therefore, a new FDR, HPP FDR is defined by considering the score for each as the maximum score for non-nested peptides per protein, where peptide scores are the maximum variant per peptide. This directly considers the best evidence for each protein as would be considered for HPP evidence, allowing for equal consideration of targets and decoys according to later filters. Traditional FDR is run on the proteins with non-zero scores. Finally, leftover FDR looks at the proteins that did not pass HPP FDR (either because they did not have sufficient HPP peptides or did not pass the score thresholds) and applies a picked FDR to these using the subset of all possible picked-pairs where neither the target protein nor the decoy protein was identified using HPP FDR. The goal here is to provide the maximum number of sensitive hints for follow up consideration.

Prioritizing future experiments

HPP-Inspector can help to prioritize future discovery and inform experiments by providing a way to curate and share results, to be further prioritized can also be further refined by an expert. To prioritize, HPP-Inspector defines three categories. The first two stem from canonical proteins with unique matches that pass global picked FDR but do not pass HPP FDR, the first category is orphans - proteins with 1+ HPP peptides and the second is hints - proteins with 0 HPP peptides. For the orphans, the question becomes finding a suitable pair for the known HPP peptide, and checking why there is no pair for the peptide. The hints might be worth following up for different reasons, and to ensure the protein has unique peptides, and if so why they are not sufficiently high quality to be considered. The final category is rejects - all other canonical proteins that pass search FDR in the dataset but are not HPP/Orphans/Hints. These are proteins that should not be prioritized from the current analysis.

Job augmentation and comparison

Any HPP-Inspector job is a valid input for another HPP-Inspector job, so it becomes straightforward to examine the combined evidence of multiple searches, including recalculation of novel proteins and application of all global FDRs in the combined input. This allows jobs to be run in an incremental manner, without reconsidering computationally expensive steps in the workflow.

Additionally, there is a way to both compare a current job, J, to another job, J', to understand how new inputs change global protein calls. Five values are output, 1. the total number of non-nested peptides in the union of the current search and the comparison, 2. the number of non-nested peptides in the union of the current search and the comparison passing HPP-FDR, 3. the number of additional non-nested peptides found from the reference and then 4. the number of non-nested peptides that are just in the reference and 5. The number of non-nested peptides only in the current job.

Workflow result views

Once the workflow is finished, the main results are presented as three interactive tables to show the evidence provided at the protein, peptide, and PSM level. For any PSM evidence, outputs are recorded as Universal Spectrum Identifiers (or USIs). This means that all evidence can be readily shared outside of these result views for other users to examine - of course the result view itself can also be shared - but having the USI provides a community-specific and publishable link to PSM-level evidence in the workflow.

The entry point is the proteins view, where all proteins in the dataset that are found to have HPP-compliant evidence can be found. For each protein, the HPP-compliant evidence is in three categories - 1. peptides that map by sequence, 2. peptides that map by sequence and match a synthetic by sequence, and 3. peptides that map by sequence and have a cosine value to a synthetic that is above a certain threshold. For each of these categories, there are an additional three characteristics - peptides that are unique to the new dataset, peptides that are unique to the library, and the union of peptides that are in the new dataset and in the library. Also, the count of HPP compliant proteins is given for each category - if for example the new dataset provided a novel sequence, it must not be a subset or a superset of what is already in the library for it to count towards calling the protein. This allows for a quick filter to see what proteins new inputs to call.

Once a protein of interest is found, there is a link from each accession to the peptides view to see the evidence for the variants that support that protein. For each sequence, statistics are shown such as how many PSMs support that sequence, what synthetics are provided for the variants, and what the best cosine between any synthetic with that variants and the input PSMs. Additionally, both the PSM for the variant with the highest database score and cosine score are shown.

Reanalysis Approach

To demonstrate the HPP-Inspector workflow, a diverse array of datasets was reanalyzed. For each dataset, spectra were searched with MS-GF+ database search against the UniProt human reference proteome with isoforms(UniProt Consortium, 2021) as well as contaminants, such as porcine trypsin. Variable modifications included in all searches were oxidation (M,+15.995), N-term acetylation (+42.011), N-term carbamylation (+43.006), Pyro-glu (Q,-17.-27), and deamidation (NQ, +0.984), and carbamidomethylation (C,+57.021) as a fixed modification. For TMT10 datasets[37], the TMT mass tag (+229.163) on both N-term and lysine, for phosphorylation datasets, Phosphorylation (STY, +79.966) is included, and for SILAC datasets heavy modifications on lysine and arginine are included. FDR was set to 1% at the spectrum level, length-specific peptide level, and protein-level for each search.

Running MassIVE-KB libraries through HPP-Inspector

Apart from an mzTab input, HPP-Inspector supports MassIVE-KB collections natively from workflow results views. To run an HPP-Inspector job with MassIVE-KB inputs, any output collection of precursors from the subheader "View Library", for example the Library Variants Ambiguity Filtered which contains all precursors in MassIVE-KB from spectra have one a single PSM that maps to that precursor and passes FDR(https://massive.ucsd.edu/ProteoSAFe/result.jsp?task=e33a302ea7e94422bf2b122260d22cc6&view=ambiguity_library_view_split). The collection can be downloaded by clicking the blue "Download" button in the header. This downloaded TSV can then be input to HPPInspector. After running this HPP-Inspector job, the resulting job can later be used as a reference when looking at how new datasets can be added to community knowledgebases (see Combinations and comparisons of multiple results sets.)

MassIVE-KB synthetic spectral libraries are also supported. These can be found for the MassIVE-KB synthetic spectrum jobs which include both synthetic peptides as well as recombinant proteins (see Matching to spectra of synthetic peptides). This input mgf can be found in the “MGF Library” link in the sub-header for “Downloads” for any MassIVE-KB synthetic job

Quality assessment of Peptide-Spectrum Matches (PSMs)

The quality of peptide-spectrum matches (PSMs) used to support protein identifications was assessed in two main ways: (a) using statistics measuring how well the peptide matches the spectrum and (b) using modified cosines to measure how well input spectra match to spectra of synthetic peptides of the same amino acid sequence. The evaluation of the quality of PSMs was based i) on how well the assigned peptide sequence explains the peaks in the spectrum (explained intensity) and ii) on the number of spectrum peaks matching to b/y ions of the assigned peptide sequence (#breaks).

Several spectrum filters were applied prior to explained intensity calculations to remove peaks that are not directly informative about the sequence of the peptide assigned to the spectrum. First, precursor peaks were removed from the spectrum, which are not informative about PSM quality since these would match any peptide of the same precursor mass (a filter that was already strictly applied by the database search algorithm). Precursor isotopes (primarily 13C isotopes) were also removed from each spectrum, as well as commonly occurring precursor neutral losses: losses of H_2O and NH_3 were removed from all spectra, loss of CH_4OS (i.e., -64 Da) from precursors of PSMs containing oxidized Methionine, and losses of H_3PO_4 and M-H_3PO_4-H_2O from precursors of PSMs containing phosphorylated residues. Second, if the PSM is annotated with either TMT[37] or iTRAQ[38]) isobaric tags then the peaks corresponding to reporter ions (including peaks for the unfragmented reporter group) are removed from the spectrum. Third, immonium ions [http://www.matrixscience.com/help/fragmentation_help.html] are also removed since these are only informative about the amino acid composition rather than the order of the amino acids on the peptide sequence. Fourth, low intensity "background noise" peaks were filtered from each spectrum using a window filter, where a peak of mass M is retained if and only if it ranks in the top *K* highest-intensity peaks out of all peaks with mass in the [M-50, M+50] interval (K=8 was used for the results presented in the main text).

Filtered spectrum peaks were then annotated using the standard ion types for collisional dissociation[39] b/y ions and their 13C isotopes and H_2O/NH_3 losses, as well as a ions. For peptides containing phosphorylated residues, the additional neutral losses of neutral loss for both H_3PO_4-H_2O and H_3PO_4-H_2O were also considered. Fragment charge states were considered from 1 up to the precursor charge state.

To further improve the robustness of the explained intensity statistic to the rare presence of very few high intensity unexplained peaks (likely caused by poorly understood sequence-specific fragmentation events), the top E highest-intensity unexplained isotopic envelopes were also removed from each spectrum. Isotopic envelopes of charge z were defined as sets of peaks whose masses differ by $1/z$, and envelope intensity was defined as the summed intensity of all peaks in the envelope. E=2 was used for the results presented in the main text.

Explained intensity (EI) in a processed spectrum S is then calculated by dividing the intensities of peaks annotated as peptide fragments by the total intensity of all peaks in S. A minimum EI of 40% was required for a PSM to be considered high quality. Additionally, the number of peptide sequence fragmentation points (#breaks) supported by either a b-ion or a y-ion was also considered, and a minimum of 5 breaks was required for the results presented in the main text.

Matching to spectra of synthetic peptides

An orthogonal way of evaluating the quality of PSMs is to assess their similarity to experimentally-acquired spectra of synthetic peptides of the same amino acid sequence[36]. If spectra of synthetic peptides are input to HPPInspector, these are also processed as described in the previous section and are matched to processed spectra of input PSMs using a modified “annotated cosine” function computing the normalized dot product[40] only between (i) peaks annotated to peptide ions in the spectrum of the synthetic peptide and (ii) peaks at corresponding masses in the input spectrum to be matched to the spectrum of the synthetic peptide.

Annotated cosines reduce the impact of interfering unexplained peaks by focusing the spectrum matching on the consistency of fragmentation patters between a PSM and the spectrum of the matching synthetic peptide (I/Isoleucine and L/Leucine were considered indistinguishable for all sequence comparisons). Two kinds of sequence matches were considered for the calculation of annotation cosines: (a) cases where the input PSM and the spectrum of the synthetic peptide have the exact same peptidoform (i.e., the same sequence with the same modifications on the same sites) and (b) cases of input PSMs with modifications that were matched to spectra of synthetic peptides of the exact same amino acid sequence but without any modifications. In the latter case the ion masses in the modified input PSM were properly shifted when matching them to the corresponding ion masses in the spectrum of the unmodified synthetic peptide.

Results in the main text used version 2.0.15 of the synthetic peptides build of the MassIVE-KB spectral library[4], as available at https://massive.ucsd.edu/ProteoSAFe/static/massive-kb-libraries.jsp. This build contains spectra of 1,662,275 distinct precursors whose spectra were collected from data acquired in the ProteomeTools[5] and BioPlex[6] (bait proteins only) projects.

Finally, PSMs passed as input to HPPInspector were considered to be reliable enough to support peptide identification only if they passed all the filters described above and either (a) had explained intensity of at least 40% or (b) had an annotated cosine of at least 0.5 to a spectrum of a synthetic peptide.

Key fields in output results tables

1. Explained intensity: the explained intensity (EI) for a peptide sequence is defined as the maximum PSM EI for PSMs of all modified or unmodified variants of the same sequence *P*. If at least one PSM for the peptide sequence has 5+ #breaks (i.e, it can support a high

quality peptide identification) then EI is defined as the maximum EI over all PSMs for *P* with #breaks $\geq T$ (T was set to 5 here), otherwise it is the defined as the maximum EI over all PSMs for *P*.

2. #breaks: number of distinct peptide fragmentation points (e.g., "backbone breaks") for peptide *P* supported by at least one b-ion and/or one y-ion with fragment charge <= precursor charge for the PSM with highest explained intensity (EI). If there is at least one PSM for *P* with #breaks$\geq T$ (T was set to 5 here), then only PSMs for *P* with T or more breaks are considered for maximum EI, otherwise all PSMs for *P* are considered for maximum EI.

3. Cosine to synthetics: the annotated cosine reported for a peptide sequence *P* and a spectrum *S* of a synthetic peptide for *P* is defined as the maximum annotated cosine between *S* and all PSMs for all modified or unmodified variants of *P*.

4. Precursor score: the precursor score for a peptide sequence P is defined as the maximum PSM score for PSMs of all modified or unmodified variants of the same sequence. As with EI, if at least one PSM for the peptide sequence has $\geq T$ #breaks (T was set to 5 here) then precursor score is defined as the maximum precursor score over all PSMs for P with #breaks ≥ 5, otherwise it is the defined as the maximum precursor score over all PSMs for P.

5. #Genes Mapped: given a peptide *P*, this is number of genes generating canonical protein sequences containing *P* or containing a peptide P' that differs from *P* by at most one amino acid (i.e., a Single Amino Acid Variation, or SAAV for short)

6. HPP compliant given a peptide *P*, this column indicates whether *P* is considered compliant with HPP criteria at both (a) the sequence level and (b) the spectrum level. At the sequence level, P must be at least 9 amino acids long and is required to have "#Genes Mapped" reported at exactly 1. At the spectrum level, *P* is considered to have enough experimental

evidence if at least one PSM across all variants with amino acid sequence *P* had #breaks $\geq T$ (T was set to 5 here) and either (a) had explained intensity of at least *E* (*E* was set to 40% here) or (b) had an annotated cosine of at least *C* to a spectrum of a synthetic peptide (*C* was set to 0.5 here). If both the sequence and spectrum conditions are met, this field is set to “Yes”. If the spectrum evidence does not meet the criteria the field is set to “No - failed quality thresholds”, and if the sequence level conditions are not met, the field is set to “No - 2+ SAAV protein matches”.

7. Database representative: PSM for peptide P with score equal to “Precursor score” as defined above. If there are multiple PSMs with the same maximal score, then the PSM that also has maximum explained intensity (EI) is selected as the representative. If there are still multiple PSMs with the same maximal score and maximal EI then one of these PSMs is arbitrarily selected as the representative PSM (deterministically set to the first occurrence of one of these PSMs in the input files).

8. Synthetic representative: PSM for peptide *P* with highest annotated cosine to a spectrum of synthetic peptide P (regardless of whether it passes spectrum quality thresholds). If there are no spectra of synthetic peptides with a matching sequence then this is set to “N/A” and the annotated cosine is set to -1.

Protein False Discovery Rate (FDR) methods

Protein sequences are considered using the same categorization as in Uniprot[35]: sequences are considered isoforms or canonical if the protein identifier begins with “sp|” and contains or does-not-contain (respectively) a "-" in the UniProt protein identifier. Other sequences in the UniProt reference human proteome [https://www.uniprot.org/proteomes/UP000005640] (e.g., with identifiers starting with “tr|”) were also considered as non-canonical sequences and were reported as isoforms in HPPInspector output tables. Contaminants are defined by the user using a separate fasta file which typically includes protease protein sequences and other common contaminants (e.g., human skin keratin protein sequences) known to be frequently detected in mass spectrometry samples. Target proteins are defined as all sequences in the fasta files passed as input by the user, and Decoy proteins are defined as the reversed sequences of all Target sequences (thus #targets = #decoys). When calculating protein False Discovery Rate (FDR), only canonical proteins and contaminants were considered as targets and only their reversed sequences were considered as decoys (matches to isoform sequences and their decoys were discarded since those do not support the detection of canonical proteins). The outline of the approaches used for protein FDR is as follows:

1. Traditional FDR. The FDR for the set of all proteins S with minimum protein score M is defined as the number of decoys divided by the number of targets in S. Given a desired protein FDR threshold T, this approach returns the largest-possible set S (i.e., the one with the most target proteins) defined by the minimum protein score $M*$ resulting in a protein FDR that is at most T.

2. Picked FDR[15] Given a set of scored target and decoy proteins, each target protein P defines a pair (P, decoy(P)) with its corresponding reversed-sequence decoy(P). Decoy(P) is set to the reverse sequence of P in HPP-Inspector, but protein pairs can also be defined for other approaches to derive decoy sequences from target sequences. The score of a pair (P, decoy(P)) is set to the maximum score of P and decoy(P). For the purposes of FDR estimation, a pair is considered a target match if P has the maximal score, and otherwise the pair is considered a decoy match. Similarly to the Traditional FDR approach for a desired protein FDR threshold T, the final set of targets/decoys is then defined as the largest-possible set of pairs (i.e., the one with the most target proteins) defined by the minimum pair score $M*$ resulting in a set FDR that is at most T.

HPPInspector defines 4 ways to do FDR using these two FDR approaches, with each way using a slightly different set of input variant precursors, protein scoring function, and FDR algorithm. A variant precursor $V = (p,m,z)$ is defined as a spectrum identification for amino acid sequence p with summed modification masses m and precursor charge z.

1. HPP FDR. Used to enforce FDR using only proteins with enough peptides meeting HPP criteria to be considered candidate HPP proteins.

   Input variants: Given the set of input variant precursors with “HPP compliant” = “Yes” (see definition above), the score of each distinct peptide sequence P is set to the maximum “Precursor score” (see definition above) across all variants with the same amino acid sequence *P*.

   Protein score: To find the score for each protein *P*, first define *S* as a non-nested subset of peptides mapping to P such that no peptide in *S* is contained in any other peptide in *S*. The score of a set of peptides is defined as the sum of the scores of the peptides in the set.

   The score of protein *P* is set to the score of the maximum scoring non-nested subset of peptides containing at least 2 peptides[36]. This set is constructed from the set of peptides $p_1, \ldots, p_N$ mapping to protein *P* at amino acid coordinates (s_i, e_i) with $score_i$ for each peptide p_i. This set is used to construct a graph G with one peptide node n_i with $score_i$ for each peptide p_i. Edges between peptide nodes are added to *G* for all pairs of node n_i with coordinates (s_i, e_i) and node n_j with coordinates (s_j, e_j) such that $s_i < s_j$ and $e_i < e_j$. Finally, *G* is also expanded with a source node with an edge to each of the peptide nodes in the graph and with a sink node with an edge from each of the peptide nodes in the graph. Using such a graph *G* constructed for a protein *P*, the HPP score of protein *P* is defined as the maximum scoring path in *G* across all paths containing at least 2 nodes in addition to the source and sink nodes. Since *G* is a directed acyclic graph, we used a dynamic programming algorithm to guarantee that it efficiently calculates the maximum scoring path across all possible paths in *G*. If there are no paths of the minimum required length then the protein score is set to zero.

   FDR approach: Run Traditional FDR on all proteins using the protein scores defined immediately above.

2. Leftover Picked FDR. Used to construct the set of Orphan or Hint proteins detected

with significant scores after global aggregation of input results, but still without enough evidence of supporting peptides to pass HPP FDR.

Input variants: All variant precursors that pass quality filters and uniquely map to a canonical or contaminant protein.

Protein score: To find the score for each protein P, first define all Peptide Precursors mapping uniquely to protein P, where a Peptide Precursor is a tuple (p,z) where p is the peptide amino acid sequence and z is the precursor charge state. For each Peptide Precursor, define its score as the maximum sum of all non-redundant variant precursor scores, where two variants $V_i = (p_i, m_i, z_i)$ and $V_j = (p_j, m_j, z_j)$ are considered redundant if $p_i = p_j$, $z_i = z_j$ and $|m_i - m_j| \leq 3$ Da. The set of maximum-score subset of non-redundant variant precursors can be found by constructing a graph for a set of variants mapping to each Peptide Precursor (p,z) such that each node is a variant of p with precursor charge z. Edges are added between variant precursor nodes $V_i = (p_i, m_i, z_i)$ and $V_j = (p_j, m_j, z_j)$ if $m_j - m_i > 3$ Daltons. The maximum-scoring subset of non-redundant variants is then defined as the set of nodes in a highest-scoring path in this directed acyclic graph. The score of a protein P is then set to the sum of all mapped Peptide Precursor scores.

FDR approach: Run Picked FDR on the subset of all possible protein pairs where neither the target protein nor the decoy protein were identified using HPP FDR.

3. Canonical Traditional FDR.

   Input variants: All variant precursors that pass quality filters and uniquely map to a canonical or contaminant protein.

   Protein score: Protein scores are defined in the exact same way as for Leftover Picked FDR.

   FDR approach: Run Traditional FDR on all proteins.

4. Canonical Picked FDR

Exactly the same as for “Canonical Traditional FDR”, but using the picked FDR approach instead of Traditional FDR.

Combinations and comparisons of multiple results sets

HPPInspector allows for Input search results to be analyzed jointly with a previously-constructed Reference results set in two ways: (i) comparison with the Reference set contrasts identifications in the input set with the identifications in the Reference set, and (ii) combination with the Reference set considers protein identifications if taking the union of peptide identifications in both the input and Reference sets. Importantly, FDRs are not recalculated for the union with the reference set because this set can be composed of just peptide sequences, which may not be associated with any identification scores (e.g., when comparing with peptide lists from resources that do not provide such scores or when identification scores are not comparable between the input and reference sets.) However, the Reference set may provide a previously determined protein FDR, and if it does this can be filtered independently of the Input set FDR. The reported combined-analysis results thus represent an upper-bound on the protein identifications that may become significant if HPPInspector is run using the Reference set as additional input search results (instead of being used as a separate independent Reference set)

For each protein, P, let Input-peps be the peptides from variants passing FDR in the input search results and let Reference-peps be the peptides from the Reference set. The following output results are then provided:

1. #HPP Peptides (Combined) Number of non-nested HPP peptides in the union of Input-peps and Reference-peps, where the protein P does not necessarily pass FDR thresholds in either Input or Reference. This is the upper bound on the number of peptides contributing to HPP evidence for a protein considering both Input and Reference.

2. #HPP Peptides (Combined, Pass FDR) Number of non-nested HPP peptides in the union of Input-peps and Reference-peps, where the protein P either passes FDR thresholds in Input or the protein P has supplied HPP-FDR below the cutoff in Reference.

3. #HPP Peptides (Reference Leftover FDR) Number of non-nested HPP peptides from Reference, only if the protein *P* has supplied HPP-FDR above the cutoff in Reference but has reported leftover FDR below the cutoff in Reference

4. #HPP Peptides (Reference HPP FDR) Number of non-nested HPP peptides from Reference, only if the protein *P* has supplied HPP-FDR below the cutoff in Reference.

5. #HPP Peptides (Added Only) Number of non-nested HPP peptides that are in the Added set that are not in the Reference set.

Each of the above categories is repeated for cases where (i) all peptides have matching synthetic sequences and (ii) all peptides have a PSM that matches to a synthetic with an annotated cosine above a certain threshold.

The following three use cases demonstrate how comparing input searches to References can help to better understand the contribution of the Input search to the Reference, with the first two cases being the set of merged results from the paper compared to community knowledge-bases, and the third case showing the difference between what one dataset can contribute only considering dataset FDR vs the impact of the dataset considering global FDR.

1. Compare merged results to KB results (https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=f243b8bb11dc45b2996aae4a9487d38b)

   In this use case, a set of searches all combined at a global 1% protein FDR is the Input and is compared to the MassIVE-KB representatives as the Reference to understand the upper-bound on the contribution from the Input set. In this example, the MassIVE-KB representatives have variant-level scores as well as predetermined protein FDR, which can additionally be filtered for.

2. Compare merged results to PeptideAtlas peptides (https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=0a9893634f374fce9f82904950e63ade)

In this use case, a set of searches all combined at a global 1% protein FDR is the Input and is compared to the PeptideAtlas sequences as the Reference to understand the upper-bound on the contribution from the Input set. Since only sequences are being used (as the identification scores are not compatible) there is no FDR threshold for Reference set.

3. Compare one dataset (PXD003947) to merged results reference (https://proteomics2.ucsd.edu/ProteoSAFe/result.jsp?task=41493daad358409e84259aac2fdc1922)

   In this use case, a single dataset at 1% dataset-level FDR is the Input and compared to the set of searches at a combined, pre-determined global 1% protein FDR as the Reference. This elucidates how some proteins are lost when the FDR threshold for the dataset is considered as compared to a set of searches.

Results

The UniProt/neXtProt categorization of five different levels of protein existence (PE) reflect the amount of evidence supporting the detection of gene protein products: PE 1 means "Evidence at the protein level" (18,407 proteins), PE 2 means "Evidence at the transcript level" (1135 proteins), PE 3 means "Evidence based on homology" (195 proteins), PE 4 means "Evidence based on prediction (gene models)" (13 proteins), PE 5 means "Evidence is uncertain" (609 proteins), where the numbers in parentheses indicate current the number of proteins in each category in the 2022 release of neXtProt. Proteins in categories PE 2, 3 and 4 are also referred to as Missing Proteins (MPs). Mass spectrometry project seeking to discover missing proteins typically start with the analysis of (i) protein extracts from tissues or fluids that are less-commonly analyzed or (ii) protein or peptide extracts obtained using various separation techniques seeking to improve the chances of detection of proteins that are otherwise difficult to detect. The 35 ProteomeXchange[41] datasets analyzed here and listed in Table 2.1 thus span a diverse collection of sample types, biological conditions and experimental protocols that are commonly used for searching for missing proteins, including 8 datasets contributed by official

C-HPP members or otherwise acquired for this specific purpose. Altogether, these datasets provide 202,008,485 spectra from 2.5TB of data. Since 80% of datasets (94% of all spectra) were submitted without any reusable identifications in standard open formats (i.e., mzIdentML[34] or mzTab[33]), and to obtain comparable results from a standardized search protocol across all datasets, all spectra were reanalyzed using MS-GF+[12] as described in the Methods section. All reanalysis results have been attached to the original datasets (see Supplementary Table 1 for all RMSVs/RPXDs). The resulting 51,567,767 Peptide-Spectrum Matches (PSMs) were matched to 3,492,532 distinct precursor variants (minimum of 495 to maximum of 1,696,363 per dataset) mapping to a combined 17,805 distinct non-decoy canonical proteins (maximum of 11,166 per search), accounting for nearly 87% of all canonical proteins and almost 93% of all PE 1 proteins with at least 1 unique sequence. As shown in Figure 2.1, earlier datasets contributed up to 2014 quickly accounted for the detection of over 10,000 human proteins, with additional gains being mostly slow and incremental except for two different kinds of experiments. First, C-HPP participants found significant numbers of missing proteins by analyzing less-common tissues and fluids, such as the 1026 new proteins reported in PXD003947's quest for missing proteins in human spermatozoa[11]. Second, enrichment and fractionation protocols allowed for deeper exploration of low abundance proteins even in tissues and cell lines that had been analyzed before, such as the 2,368 new proteins reported in PXD004452's exploration of multiple cell lines using extensive fractionation, various enzymes and peptide enrichment strategies[42].

But while many datasets consistently contributed detections of previously-missing proteins, it is important to consider that any analysis (or reanalysis) of increasingly larger volumes of data also creates more possibilities for false positive protein identifications (especially as the set of still-missing proteins becomes smaller and smaller). To illustrate the impact of this problem, Figure 2.1 also tracks the increase in false positive identifications across datasets by accumulating both target and decoy identifications, thus revealing that the protein-level false discovery rate for naively accumulated identifications was over 31%. The HPP-guidelines[36] for discovery of missing proteins require addressing this problem with additional criteria for the identification of both peptides and proteins. As described below, HPP-inspector's implementation of HPP guidelines criteria improved the quality of the results but also resulted in very substantial reductions in the number of proteins supported by the data, with up to 1,524 proteins lost for dataset PXD010025[43] and over 77% of all proteins lost for the secretome dataset PXD005656[44] due to overall low identification numbers leading to fewer proteins having 2+ peptides.

Rigorous analysis of peptide-level results requires careful definitions for what is meant by "peptide" identifications. The HPP-inspector workflow illustrated in Figure 2.2 uses three different categories for this purpose. First, peptidoforms are identifications where specific post-translational modifications are assigned to specific sites on the peptide sequence, such that the same sequence with the same modifications on different sites constitute different peptidoforms. Second, variants of a sequence are identifications of the same exact amino acid sequence with the same summed modification masses. In this case, a single variant of a peptide P could represent multiple peptidoforms with different site localizations of the same modifications and would also group a peptidoform with a di-oxidized amino acid together with a peptidoform of the same sequence with two singly-oxidized amino acids. Variant-level analysis is especially useful for HPP-Inspector analysis because the input search results are typically not properly controlled for errors in modification assignments (e.g., two single oxidations vs di-oxidation) and modification site false localization rates[45]. Finally, peptide identifications of a sequence group together all modified variants and all precursor charge states of the same amino acid sequence. Peptidoforms and variants with different precursor charges are referred to as different peptidoform and variant precursors, respectively. All categories group together identifications of sequences differing only by Isoleucine (I) or Leucine (L) since these cannot be distinguished in almost all commonly used mass spectrometry data acquisition protocols.

While additional controls are not necessary to guarantee a global 1% PSM-level FDR (the union of PSMs aggregates both target and decoy matches so the FDR stays the same), it is still necessary to recalculate precursor-level FDR when merging results from multiple searches. Since different peptidoforms of the same amino acid sequence P with the same modifications do not contribute independent evidence for the observation of P, HPP-Inspector imposes precursor-level FDR at the level of sequence variants (thus avoiding potential ambiguities with modification masses or site localization). As required by the HPP guidelines[36], peptide identifications must also meet two additional criteria beyond FDR: guideline 5a states that peptide identifications should be made using “high signal-to-noise ratio and clearly annotated spectra” and guideline 9 states that peptide sequences should be at least 9 amino acids long and match only a single protein sequence even when considering Single Amino Acid Variations (SAAVs). HPP-Inspector implements guideline 5a using quality filters requiring that variants be represented by at least one PSM with at least 40% explained intensity (EI) and containing enough fragment ions (either b or y ions) to support identification of at least 5 peptide fragmentation points (#breaks, see supplementary information for additional details). Starting from 3,492,532 variants passing FDR in the union of search results for all 35 datasets, HPP-Inspector quality filters removed 171,855 variants (5% of all) from consideration. After imposing quality filters, HPP-Inspector then imposes HPP filters implementing guideline 9, which further remove 497,703 variants (14% of all) from consideration, mostly because of shared peptide sequences matching to 2 or more proteins. Figure 2.3a shows the filtration impact of each of these filters using an UpSet plot[46] showing the counts for each category where variants do/do-not meet each of the 4 criteria. A total of 2,838,365 variants (81% of all) was thus retained after applying all filters.

The HPP guidelines also recommend that whenever possible, the reported variant PSMs should also be matched to spectra of synthetic peptides to further establish the correctness of novel discoveries. As such, HPP-Inspector also provides the option for users to input a spectral library containing spectra from synthetic peptides (see Figure 2.2). If such a library is available, HPP-Inspector matches the fragmentation of the annotated ion peaks in the spectral library to the fragmentation pattern of the same peaks in the corresponding identification in the input search results (see supplementary information for details). The current release of the MassIVE-KB spectral library[4] (v2, available at https://massive.ucsd.edu/ProteoSAFe/static/

massive-kb-libraries.jsp) containing reference spectra for 1,666,936 precursors was used for the results reported here. HPP-Inspector considers two types of variant matching to determine if a PSM from a synthetic peptide is eligible for matching to a reported PSM identification. Type 1 matches consider only cases when the variant string is the same for both the reported and synthetic peptide identifications, thus requiring that both PSMs be for the same peptide sequence with the same summed modification masses. To increase the utility of spectra of synthetic peptides for matching to PSMs of modified peptides, HPP-Inspector also considers Type 2 matches where an identification PSM with amino acid sequence P is matched to all library variants of the same sequence P regardless of modifications. If an identification PSM can be evaluated using both Type 1 or Type 2 matches then the highest match cosine of both types is reported. As shown in Figure 2.3b, HPP-Inspector finds that for high-quality (according to quality filters) Type 1 matches the reported identification is confirmed for nearly all cases (98.6%). As expected, the percentage of confirmed identifications drops slightly to 92.4% of cases for most high-quality Type 2 matches. The major exception observed to the latter was for reported identifications containing TMT modifications, where the changes in fragmentation patterns introduced by the N-terminal labeling groups resulted in only 35.5% of cases being confirmed by high-quality Type 2 matches. Using spectra of synthetic peptides for analysis of low-quality identifications showed that significant fractions of low quality PSMs still matched the fragmentation in the spectra of the corresponding synthetic peptide. Since matching to spectra of synthetic peptides is an orthogonal assessment of PSM quality, HPP-Inspector allows for PSMs with low EI or low #breaks with high cosine to spectra of synthetic peptides to also be considered as evidence for novel discoveries (this can be enabled or disabled using a configurable option in the workflow input form).

A natural response to the question of whether a gene protein product *P* has ever been observed would be to consider whether *P* has ever been reported as detected in any published proteomics study. However, considering the union of proteomics studies aggregates false positive as well as true positive discoveries, and this usually results in unacceptably high false discovery rates (as shown by the 32% FDR in Fig 1 for the set of proteins from just 35 datasets). Figure 2.4 shows how the various levels of protein, peptide and PSM filters affect both the FDR and the reported number of significant proteins. First, requiring that proteins be identified using only peptides that map uniquely to protein products from a single gene eliminates 599 target proteins identified by parsimony using shared peptide sequences mapping to protein products from 2+ genes. Second, applying the PSM quality filters above eliminates an additional 233 target proteins from consideration, but also significantly reduces FDR from 31% to 22% (as expected since decoy PSMs are expected to be of lower quality than true target PSMs). Third, requiring that proteins be matched by HPP peptide sequences (i.e., at least 9 amino acids long and mapping to products from only one gene even if considering one SAAV) substantially reduces the set of discovered proteins – 386 proteins lost when requiring 1+ HPP peptide and an additional 955 proteins lost when requiring 2+ non-nested HPP peptides (see supplementary information for details).

Two examples illustrate how HPP criteria affect the set of resulting proteins. In one case, PE 1 protein sp|Q2VWA4|SKOR2_HUMAN is matched by 7 distinct variants (each of which individually satisfying HPP criteria), with the variants mapping to amino acids 368-378, 425-448, and 801-818 on the protein sequence. However, regions 368-378 and 801-818 each have only a single low scoring PSM for the variant, whereas there are 5 nested, but much higher scoring variants for the region 425-448 – this greatly increases the protein-level score for parsimonious or canonical protein analyses but yields only a modest protein score when enforcing HPP criteria. In another case, the PE 2 protein sp|Q6NXN4|D19P1_HUMAN is matched by 2 distinct variants that satisfy HPP-criteria, and then also matched by TKM[+15.994915]GLYYSYFK/2. However, this latter variant also maps to the PE 1 protein sp|Q6NUT2|D19L2_HUMAN with sequence TEM[+15.994915]GLYYSYFK/2 and a E→K (-0.9414 Da) single amino acid variation (SAAV) on the second amino acid on the peptide sequence. Even though the PSMs for this identification clearly match the TK version of the peptide much better than the TE sequence, this is a known SAAV reported in neXtProt and is a case where the codons for the two amino acids (K and E) differ by only one nucleotide, thus also constituting a Single Nucleotide Polymorphism (SNP). Combining the SAAV information with the observation that PE 1 protein sp|Q6NUT2|D19L2_HUMAN is also supported by over 200 variants (including 141 HPP-compliant variants), further supports the preferred interpretation of this peptide as a less-surprising SAAV on PE 1 protein sp|Q6NUT2|D19L2_HUMAN instead of allowing it to support a surprising novel discovery of PE 2 protein sp|Q6NXN4|D19P1_HUMAN.

Although these HPP requirements lead to a significant drop in sensitivity, the need for enforcing these requirements is also supported by their very significant impact in reducing FDR from 22% to 7%. Finally, enforcing 1% protein level FDR eliminates an additional 256 proteins to yield the final set of 15,376 proteins passing all quality and FDR requirements (an aggregate loss of 2,429 proteins from the initial 17,805 reported in the 35 input datasets). The additional requirement in the HPP guidelines to match the search PSMs to spectra of synthetic peptides is only possible to evaluate for 13,876 proteins (90.2% of all passing all previous quality filters), even when using the MassIVE-KB spectral library constructed from multiple datasets containing data from synthetic peptide and protein sequences. That said, the robustness of the HPP and quality filters implemented in HPP-Inspector is reinforced by the synthetics-confirmed identification 13,767 proteins, corresponding to 99.2% of all with PSMs for 2+ HPP peptides that can be matched to spectra of synthetic peptide sequences (with most of the unconfirmed cases following the same patterns discussed above for the analysis of precursor identifications).

The low number of only 2 new missing proteins (MPs, i.e., proteins in categories PE 2, 3 and 4) reported in the HPP-Inspector analysis of the 35 datasets is not surprising because all datasets were released before this year and their data has already been integrated in the updated MassIVE-KB build integrated in the 2022 neXtProt release. However, the contributions of each dataset to the discovery of novel proteins is best assessed by the number of MPs detected in the datasets at the time when the datasets were released. HPP-Inspector evaluates these contributions by also reporting results for historical versions of neXtProt, staring with the 2014 release. As thus shown in Supplementary Table 1, we can see that several datasets contributed data for the discovery of dozens of novel MPs in the years when they were released, with up to 63 new MPs first detected in 2017 in dataset PXD004452 (which considered various deep fractionation protocols for the analysis of cell lines).

However, these two new MP are both due to isoforms that are potentially from read-through transcripts, but which are not recorded as read-through. This is an issue when calling the MP Q9Y4R7 from gene TTLL3, as the isoform P59998-2 is a readthrough-protein of ARPC4-TTLL3, however is only considered a protein from the gene ARPC4, and since all sequences unique to Q9Y4R7 to also map to P59998-2, these sequences appear to be not HPP compliant as they map to multiple genes. A similar situation occurs with MP E9PB15, from gene PTGES3L, when there is an isoform Q9BTE6-2 which is noted to be a read through from PTGES3L-AARSD1, noted on the Uniprot website with manual curation[35]. In both cases, since these isoforms contain read-through transcripts with other genes, these proteins would never be called unless the gene is updated in Uniprot, or the guidelines do not consider isoforms in determining uniqueness.

Since the novelty of discoveries in new datasets depends on the peptide and protein observations already recorded in existing knowledge bases, it is important for HPP-Inspector to be able to both compare and consider the integration of new identifications with existing knowledge base identifications. This is implemented by allowing users to submit a collection of reference knowledge base results, which can be provided either as a spectral library (e.g., in MassIVE-KB format) or as a simple list of peptide sequences (see supplementary information for details and use cases). When comparing to the original release of MassIVE-KB, the union of new results from the 35 datasets would contribute new evidence for the detection of 662 additional proteins. Conversely, MassIVE-KB already contained evidence for 847 proteins not observed in the 35 datasets so HPP-Inspector also reports that the combination of search results could generate a combined set of 16,223 proteins using peptide identifications from both older and more recent datasets. HPP-Inspector also reports results for three additional categories of protein identifications that can be used to help prioritize future experiments seeking to reveal additional evidence for the identification of missing proteins. Starting from the set of proteins with unique peptide matches that pass canonical picked FDR but do not pass HPP FDR (see supplementary information for details), HPP-Inspector categorizes a protein P in this subset as an Orphan if it is matched by at least one HPP peptide and as a Hint if it is matched by no HPP peptides. In both cases, the detection of the proteins in specific samples from specific proteomics studies, tissues or cell lines provides partial evidence of protein expression (or at least improved detectability) under those conditions, which could be used to design additional experiments on the same type of biological samples to hopefully acquire the additional missing evidence to establish the discovery of the novel proteins. Even when considering the new results from the 35 datasets in addition to proteins already identified in the original release of MassIVE-KB, HPP-Inspector reports 670 Orphan proteins (including 70 MPs) and 282 Hint proteins (including 25 MPs), thus providing partial evidence and direction for follow up research with the potential to identify 952 additional proteins, including potentially identifying 95 additional novel MPs.

Discussion

As the community gets closer to finding mass spectrometry evidence for the entire translated proteome, HPP-Inspector allows for individual contributors to understand the impact of their experiments in the context of community-scale knowledge. Many of the 35 datasets came from different labs and represent different sample types, labeling protocols, etc. but all can be analyzed in the same workflow. Further, while HPP-Inspector will inform the user of what is likely sufficient for community scale knowledgebases, it will also provide directions for follow up, in the forms of orphan and hints as well as show which precursors and proteins have matching synthetics. And even if the quality is sufficient for the dataset-level analysis, the workflow can be used to compare current results at the peptide level to community libraries.

Acknowledgements

Chapter 2, in part is currently being prepared for submission for publication of the material. Pullman, Benjamin S.; Wertz, Julie; Carver, Jeremy; Bandeira, Nuno. The dissertation author was the primary investigator and author of this material.

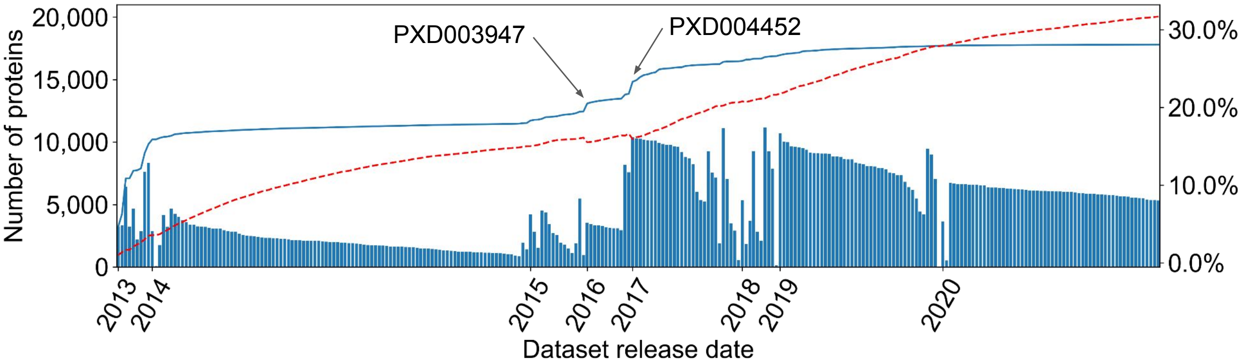


Figure 2.1. Histogram shows the number of target proteins identified per search in each dataset (blue bars) and the accumulated target proteins (blue line) for all searches of the 35 datasets, adding searches in chronological order of the date of dataset submission to ProteomeExchange. As new searches are addeded to the aggregate set of results, the protein-level False Discovery Rate (FDR) steadily increases (red dotted line) at a rate higher than new protein identifications since targets proteins generally coalesce (i.e., same proteins identified across many datasets) and decoys do not (or do so at lower rates). The unfiltered acceptance of all results derived from all datasets thus results in an unacceptably high 32% protein-level FDR.

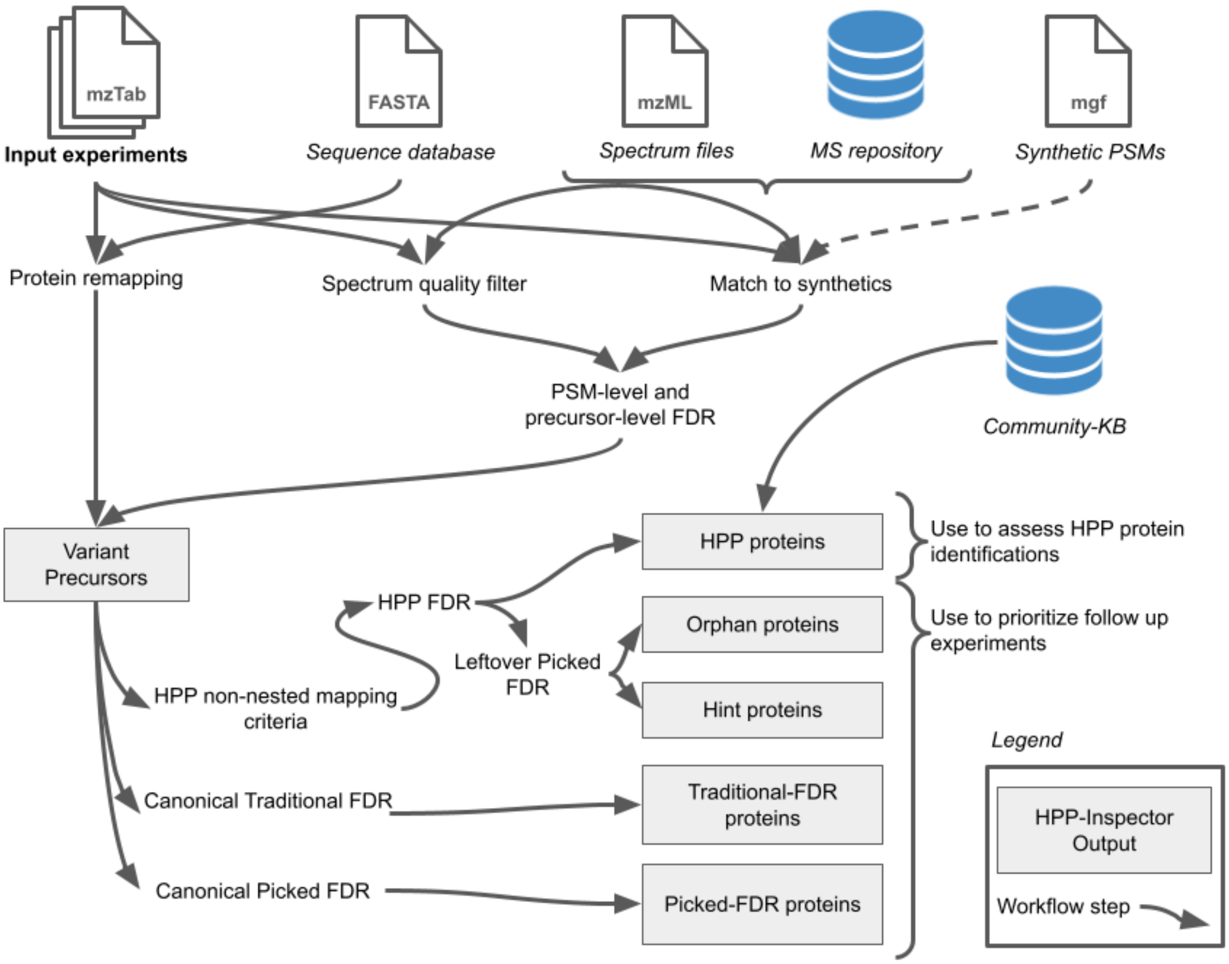
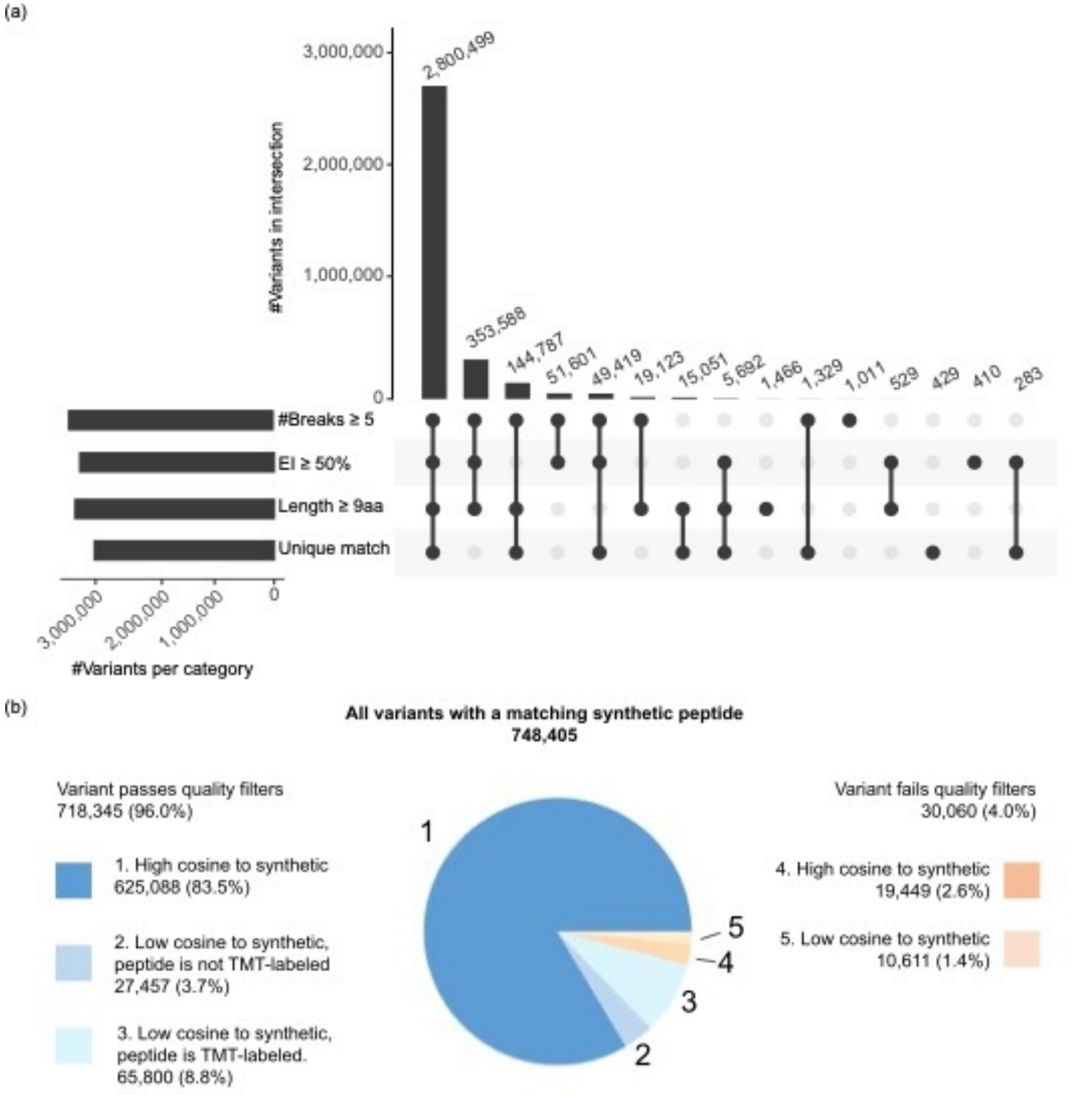


Figure 2.2. The overall process of HPP-Inspector from coalescing input PSMs, to applying multiple levels of FDR, to the HPP-compliant output. Each step can be scrutinized by the user through online tables and all PSMs, including their synthetic matches, can be shared through a USI[47]. Additionally, the HPP-Inspector workflow outputs proteins that are not HPP-compliant but can help inform future experiments about where unseen or missing proteins might be found.

Figure 2.3. (A) UpSet[46] plot for all variant precursors to assess which of the 3,492,532 input precursors are HPP-compliant according to the four precursor-level categories of 1) Only mapping to a single protein including single amino acid variants (SAAVs), 2) Having a PSM for the variant precursor with more than 40% explained intensity, 3) 9+ amino acids in length, and 4) Matching 5+ backbone breaks in the peptide. The bottom plot shows all the different possible intersections between the categories, showing notably that around 78% of all variant precursors satisfy all conditions. (B) shows all precursor variants have a match to a synthetic PSM, breaking the matches down into two categories, precursor variants that pass PSM-quality filters and those that do not. Among those that do, nearly all have high cosine to synthetics (1), except for a small number where the TMT labelling caused the match to be lower quality (2) and about half of that which have differing fragmentation for other reasons (3). A small number of spectra matched synthetic sequences with high cosine but were lower quality (4) and others both considered low quality and confirmed to have a low cosine to synthetics, implying some difference in the fragmentation compared to what is expected for that precursor (5).



		Precursor passes quality filters			**Precursor fails quality filters**		
Sequence match	**Type**	*High cosine*	*Low cosine*	*% pass*	*High cosine*	*Low cosine*	*% pass*
Variant	No TMT	301819	4268	98.6%	3242	804	80.1%
Peptide	No TMT	286992	23716	92.4%	15813	8831	64.2%
Peptide	TMT	36277	65800	35.5%	224	1057	17.5%

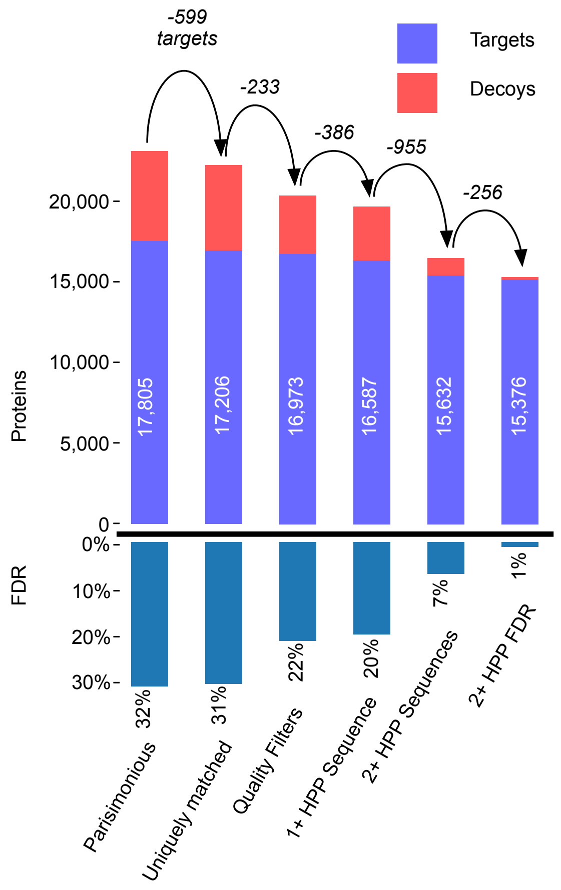


Figure 2.4. The first bar shows the 17,805 proteins from the union of all search jobs, each at 1% parsimonious protein FDR, leading to a 31% FDR. The second bar still considers per search FDR, but only considers uniquely matched proteins, still at 30% FDR but losing 599 targets. The next bar imposes spectrum quality filters, still maintaining 16,973 targets, but reducing the FDR to 21%. Requiring 1+ and 2+ HPP peptides further limits the number of targets, but ensures all targets meet HPP criteria, however the FDR is still 7%. Finally, imposing an HPP-FDR of 1% filters the set of proteins down to 15,376 total target proteins, that all pass HPP criteria

Table 2.1. The 35 datasets are sorted chronologically by submission date and by the number of proteins identified from each dataset using parsimonious, unique, and HPP False Discovery Rate (FDR). Dataset-level FDR is recalculated over all searches of the subsets of data from each dataset (e.g., separate searches for different tissues or cell lines), such that the reported numbers represent 1% dataset-level FDR.

PX Dataset Identifier	Year	#Spectra	#PSMs	#Variant Precursors	#Proteins Parsimony	#Proteins unique peptides	#Proteins 2+ HPP peptides
PXD000442	2013	448,336	121,070	27,872	3,361	3,256	2,368
PXD000447	2013	354,332	82,901	23,148	3,254	3,146	2,258
PXD000443	2013	469,536	232,550	88,793	6,513	6,359	5,076
PXD000449	2013	969,814	293,870	52,984	4,802	4,724	3,350
PXD000263	2013	165,699	31,598	15,466	2,895	2,804	1,863
PXD000529	2013	1,546,542	568,384	135,196	7,700	7,544	6,392
PXD000533	2013	2,167,835	843,006	175,485	8,418	8,264	7,115
PXD000427	2014	394,324	100,431	21,154	2,916	2,809	1,687
PXD000603	2014	167,128	37,753	7,942	1,782	1,675	863
PXD000900	2014	14,292,385	3,062,728	349,881	6,211	6,151	5,705
PXD000547	2014	242,723	41,438	10,389	1,429	1,361	891
PXD000548	2014	333,713	45,624	14,649	1,976	1,904	1,296
PXD000754	2015	519,326	176,289	37,125	4,266	4,129	2,957
PXD001694	2015	2,228,403	111,604	26,158	3,106	3,009	1,807
PXD002255	2015	4,161,568	960,666	153,185	7,446	7,332	6,190
PXD001933	2015	530,308	189,295	47,231	5,534	5,386	4,160
PXD002428	2015	1,536,403	8,066	2,916	873	834	339
PXD003947	2016	2,568,210	1,077,954	161,284	5,021	5,009	4,147
PXD004785	2016	4,216,141	1,879,600	179,645	8,516	8,417	7,514
PXD004452	2017	22,284,971	8,250,139	1,376,655	13,571	13,415	12,788
PXD006798	2017	737,902	59,024	10,800	1,916	1,835	908
PXD006833	2017	8,370,043	2,758,725	380,278	11,230	11,035	9,945
PXD006465	2017	4,107,526	599,858	160,036	7,201	7,098	5,943
PXD006557	2017	796,492	20,694	1,844	558	511	137
PXD010025	2018	2,111,875	258,145	79,027	5,392	5,257	3,821
PXD009646	2018	1,773,164	45,100	17,883	1,878	1,803	1,286
PXD010142	2018	1,483,231	887,238	41,639	3,739	3,643	2,806
PXD010093	2018	946,983	70,912	31,924	3,250	3,215	2,250
PXD009737	2018	5,766,845	1,178,891	503,069	11,726	11,565	10,311
PXD005656	2018	94,891	1,923	491	150	135	35
PXD010154	2019	76,177,914	20,074,287	1,646,516	14,313	14,193	13,571
PXD004092	2019	7,104,983	2,599,552	285,481	9,852	9,760	8,833
PXD014083	2019	3,507,193	461,004	122,868	7,116	6,944	5,817
PXD014300	2020	374,729	144,935	33,053	3,894	3,769	2,566
PXD016999	2020	29,057,017	4,292,513	228,499	9,000	8,456	8,278

Chapter 3

StrataCluster: a new approach to tandem mass spectrometry clustering for accurate estimation of the missing human proteome

Introduction

Mass spectrometry (MS) is the main technology for high throughput analysis of proteomics samples[48], allowing for high throughput analysis of proteins and post-translational modifications of cancer samples[49], protein biomarkers[50], and drug targets[51]. As such, there have been systematic efforts in the proteomics community to create a mass spectrometry-based blueprint of the human proteome to understand human biology and disease[8][10][26]. The increasing quality of MS data and algorithms have enabled many impactful proteomics results and the recently released blueprint of the human proteome, but have also led to the incorrect perception that the problem of identification of tandem mass (MS/MS) spectra is mostly solved.

In difference from this, algorithms for clustering of MS/MS spectra have long indicated that large fractions of MS data remain unidentified, but high error rates in mixing (incorrect grouping of different peptides) and splitting (incorrect separation of spectra from the same peptide) have significantly complicated the task of quantifying the portion of the human proteome that still eludes current identification algorithms.

MS/MS clustering algorithms in mass spectrometry have traditionally been designed to maintain cluster quality (all clusters should contain spectra which can be labeled as the same peptide) with the purpose for speeding up and improving the quality of searches while lowering cluster redundancy somewhat, something which the algorithms have been shown to do accurately and quickly[52][53][54][55][56][57]. We propose a novel formulation of the MS/MS spectra clustering problem eliminating the ubiquitous assumption that each spectrum is generated from a single peptide, and show that the StrataCluster algorithm results in clusters with $>$ 15-fold less mixing and $>$ 17-fold less splitting than current state of the art approaches.

While it might seem that current algorithms can give decent lower bounds by changing parameters to reduce cluster quality in exchange for fewer, and more representative clusters, we found this not to be the case. It then becomes necessary to reconsider the approaches - to attach equal emphasis on creating high quality, representative clusters (measured as mixing, and defined it as the number of identifications in a given cluster) and at the same time, ensuring that the number of clusters generated is not significantly higher than the number of underlying molecules the spectra represent (splitting and it is defined as the number of overall identifications as compared to the number of overall identified clusters).

StrataCluster results thus reveal that as much as 94% of the human proteome is currently missed by typical data analysis protocols, including 69% of medium-abundance peptides detected in multiple human tissues. Probing into this “missing proteome´´ using open-modification search and spectral alignment algorithms further reveals two separate categories of missing identifications.

First, the “gray proteome´´ (38% of all clusters) reveals hundreds of thousands of post-translationally modified peptides that are routinely missed by typical search protocols and are only partially recovered by open-modification searches. Accounting for even more clusters than are typically identified (only 30%), we show that these define diverse collections of “peptide families´´ aggregating many modified peptide variants, including many hypermodified peptides found to be differentially-abundant across human tissues. Second, this analysis also revealed a “dark proteome´´ (32% of all clusters) that are not significantly matched to any sequences in the reference human proteome, even though over 130,000 of these also form peptide families with variants that also differ by known modification masses.

Finally, clustering results across the deep characterization of 29 human tissues reveals very similar patterns of variation for both commonly-identified and missing-proteome peptides, including tens of thousands of tissue and sample-specific hyperexpressed peptides, where only 36% are commonly identified and the remaining split into 35% gray proteome and 28% dark proteome peptides. Overall, the analysis enabled by StrataCluster's novel clustering approach indicates that improved algorithms for analysis of the gray proteome have the potential to double the number of medium abundance peptides identified in proteomics experiments, while also showing that new approaches are necessary for analysis of the dark proteome (which still remains almost as large as the current commonly-identified proteome).

In addition, StrataCluster's detection of patterns of variation across tissues (including sample-specific hyperexpression) and construction of peptide families detecting relationships between identified and/or unidentified peptides can be used to both prioritize and facilitate the follow up analyses that will be necessary for advancing the elucidation of the gray and dark human proteomes.

Methods

Dissimilarity edges

Dissimilarity edges model spectrum-spectrum matches as potential containments instead of as full spectrum relationships. The edge is defined by considers the top k peaks ranked by intensity of one spectrum being present in all peaks of another (see Dissimilarity Edges Supplement). Being a containment relationship, dissimilarity edges are directed as the top peaks of one spectrum can appear in another, though the converse might not be true. If there are two edges, one in each direction, that is referred to as a bidirectional dissimilarity edge, representing cases where the two spectra likely are from same peptide (though cosine similarity is used to confirm this). For bidirectional dissimilarity edges, most spectrum-spectrum pairs that share the same peptide are captured. However, many spectrum-spectrum pairs can only be modeled as asymmetric dissimilarity edges indicating that only considering exact matches these relationships would be either mischaracterized or missed. If there is only one dissimilarity edge between two spectra, that is called a unidirectional dissimilarity edge and this represents cases where one spectrum is a subset of another, potentially either due to co-elution or incomplete fragmentation. Beyond the recovered spectrum-spectrum pairs that might have been lost when only considering spectra that match exactly, dissimilarity edges allow for some algorithmic speedups, allowing us to use a hashing technique (see Dissimilarity Graph SI) to find these edges, mitigating the need for many costly spectrum-spectrum comparisons. This creates a reduction of comparisons by on average 1000-fold, for bins with many precursors.

Clustering algorithm

To ensure that the output clusters are as non-redundant as possible, without sacrificing sensitivity, spectra are clustered in a way that maintains a stratification within the cluster. First, dissimilarity edges between all spectra within a given parent mass tolerance are constructed, using a hashing process to speedup construction of the graph. Within each connected component, there are three steps, first considering all spectra that have only bidirectional dissimilarity edges and are at the same parent mass, then relaxing the constraint on parent mass, allowing matches that are 1 isotope away, and finally allowing for unidirectional dissimilarity edges to be considered.

The first stage groups the many spectra that are nearly identical in the connected component. While bidirectional dissimilarity edges can be extremely precise with $< 0.1\%$ error, these edges are not sufficient to completely discriminate between two spectra at a large scale, as with hundreds of millions edges in a clustering graph, 0.1% could quickly create an abundance of mixing. To avoid this, the cosine similarity or the dot product of two normalized vectors is added as an added criterion for similarity. To begin the clustering process, the unclustered spectra in each connected component are first ordered by their best outgoing edge, based on cosine similarity, with the intention of seeding the clusters with spectra that are as similar to each other as possible. For each spectrum, form a new cluster with the two spectra of the highest cosine edge. Then repeatedly consider the spectrum that has the highest average cosine of at least theta and a bidirectional dissimilarity edge to at least 80% (since at the given cosine threshold the recall of all true edges is around 80%) of the spectra in the cluster and add them to the core of each cluster, removing them from the unclustered set of spectra until there are no spectra to add to the cluster. Finally, to find the representative of this cluster. Consider pairwise similarity between all spectra in the cluster and set the spectrum that has the highest average pairwise similarity is set to the first representative of the cluster. Repeat the process until all spectra are considered, in all connected components.

The second stage considers spectra that should have been clustered but due to different permutations in peaks and parent masses that can differ by an isotope, might have been missed. The previous stage is repeated, however instead of considering unclustered spectra, only the representatives for each cluster are considered and now allowing for one isotope difference in parent mass between representatives. Order the clusters by size and for each cluster considered, repeatedly add new clusters that contain a bidirectional edge to at least 80% to all representatives and an average cosine of at least theta by adding the new cluster representative to the set of representatives for the cluster and adding all the spectra to the set of core spectra. Repeat this until all clusters are considered.

The final stage considered the unidirectional dissimilarity edges and all the containment relationships that have not been considered, leaving many related clusters unconnected. For unidirectional dissimilarity edges, projection cosine, a measure of spectrum containment17, is used to better assess the quality of the match, even if the full spectrum cosine is not high. When considering unidirectional dissimilarity edges, the working assumption is made that larger, more homogeneous clusters are more likely to be the highest quality spectra for the peptide and should be the cluster representative. Picking the cluster representative then determines if the edge created is a subset (the spectrum with more peaks is a mixture) or a superset (the spectrum with fewer peaks is missing signal). If there is a tie, the representative is chosen using minimum KL divergence for superset edges, and maximum precursor intensity for subset edges. Order the clusters by size of the core and for each cluster, c, repeatedly add the cluster, c', connected via either any dissimilarity edge and with the highest average projection cosine to all representatives of c of at least theta', as follows: c becomes a satellite of c if the edges between c and c' are bidirectional, c' becomes a weak satellite of c if the edges between c and c' are unidirectional with c' being a subset of c, and finally c' becomes a mixture of c if the edges between c and c' are unidirectional with c' being a superset of c. Repeat until no two clusters can be merged.

Database search

Both unclustered spectra from the dataset and cluster representatives from all tools are representatives are searched using MSGF+[12] against the UniProt human reference proteome19 with isoforms and contaminants, including porcine trypsin. The results are filtered to a 1% PSM and 1% precursor level FDR. Variable modifications included in the search were oxidation, N-term acetylation, N-term carbamylation, Pyro-glu, and deamidation, and carbamidomethylation on cysteine was considered as a fixed modification. A parent mass tolerance of 0.01 Da was used and one missed cleavage was allowed. For the representative-level evaluation, the representatives output from each algorithm are identified in the same manner as the spectrum-level results.

Evaluation criteria

To consider how StrataCluster compares to other algorithms, there are four criteria to consider. The first, mixing, is the percentage of clusters that contain two or more top-scoring distinct sequences, from searching the individual spectra. A cluster with one sequence is considered pure. Distinct sequences (and not peptides) are considered since all algorithms to be evaluated generally cluster within a narrow parent mass tolerance as the goal is not to score mixing based on localization of modifications. The second, splitting, is the total number of identified clusters divided by the total number of unique cluster representatives minus 1. This assess how many extra clusters there are beyond the goal of having one cluster per molecule. The third, precursors output, is the number of precursors from cluster representatives after accounting for potential ambiguity with mass error vs deamidation and carbomylation vs acetylation. This measures overall sensitivity of the approach. Finally the fourth, adjusted unidentified precursors, is the number of potential precursors left to identify. This is calculated as the number of unidentified precursors divided by 1+splitting.

Cascade search

In order to reduce false positive identifications in searches allowing for unexpected modifications, clusters representative spectra were first searched with MSPLIT spectra library search[58] against the MassIVE-KB spectral library[4]. This search considers a smaller search space (only peptides whose spectra have been confidently identified before) and has access to the most information per peptide (i.e., peptide spectrum fragmentation is known a priori and represented in the spectral library reference spectra per peptide). In addition, MSPLIT searches are also able to identify mixture spectra containing peaks from co-eluting peptides. Second, representative spectra that are not identified by MSPLIT are searched using the same MSGF+[12] protocol described above to allow for the identification of precursors with no representatives in the spectral library or with significantly different peptide fragmentation patters (e.g., possibly due to co-elution with precursors not represented in the spectral library). Finally, only representative spectra that are not identified by MSPLIT or MSGF+ are considered for searching using MODa20 to consider the largest search space of peptides possibly containing unexpected modification masses. In difference from similar open-modification search tools21 which search all spectra against all sequences in the protein database, this MODa search considers only spectra that were not identified by MSPLIT/MSGF+ (thus reducing the chances of obtaining false positive identifications with unexpected modifications when simpler explanations are statistically significant in more restrictive searches) and only considering peptide sequences from the same proteins that were identified by MSPLIT/MSGF+ (thus reducing the space of sequences available for possible false positive identifications).

Spectral networks analysis

Spectral networks are graphs defined on sets of spectra where each node represents a spectrum in the set and edges represent the detection of significantly correlated peptide fragmentation between the spectra of the connected nodes. For the analysis reported here, each cluster defines one spectral network node corresponding to the cluster representative spectrum and each node's spectrum counts is set to the number of spectra in the corresponding cluster. As previously described[59][60], Smith-Waterman[61]-like spectral alignment algorithms are used to define the edges of a spectral network based on the detection of correlated fragmentation in spectra of related peptide sequences differing by modifications, longer/short prefixes/suffixes or amino acid polymorphism. Spectral alignment was only considered for pairs of nodes sharing at least one short de novo sequence tag of length 3 in their top 50 sequence tags. When spectral alignment was computed to define edges between nodes, it allowed for peak matches between the aligned spectra regardless of the parent mass difference and used the normalized dot product to assess the similarity of aligned spectra, as described before for the spectral networks definition of molecular families[60].

Results

Stratified clustering framework

To build a clustering algorithm designed for characterizing the dark proteome sensitively and without redundancy, it is necessary to devise a clustering framework that can categorize the range of spectra seen for a given peptide, identified or not. Clustering moves away from being designed for uniformity (i.e. only the most similar spectra) and towards a strategy that stratifies spectra into three main components, as shown in Figure 3.1A. 1. The core, which is where the high-quality spectra that define the representative of the spectra are located. 2. The satellites, which are for low quality versions of the representatives from the core and 3. The mixtures, which are like the core but instead have extra peaks, whether due to coelution or extra noise peaks. The structure maintains all spectra associated with a peptide in the same cluster, without enforcing a homogenous similarity requirement for all types of spectra for membership in clusters.

Comparison to other tools

StrataCluster performs well compared to the state of the art in clustering algorithms, MaRaCluster[56], MSCluster[52][53], spectra-cluster[54][55], and GLEAMS[57]. All these algorithms perform the task of spectral clustering using a form of agglomerative clustering. MaRaCluster also explicitly handles the problem of coeluting peptides, in both the treatment of spectrum-spectrum similarity measures as well as requiring complete-linkage when creating clusters. MSCluster , spectra-cluster, and GLEAMS all do not consider coelutions in their clustering. However, StrataCluster is the only algorithm specifically designed to find a lower bound of the number of clusters while the others are designed for the task of reducing the number of input spectra but maintaining all identifications in the output. Therefore, while all the algorithms are concerned with mixing, StrataCluster also places high emphasis on splitting.

All the algorithms are used to cluster the HCD, tryptic, and non-synthetic spectra in MSV000083508. To assess the quality of the clusters, both individual identifications for each spectrum contained in each cluster (spectrum-level) and also the identification of the representatives (representative-level) are considered. Comparing StrataCluster to other MS/MS clustering algorithms, mixing and splitting are both significantly lower than the other tools, especially with regards to splitting, as seen in Figure 3.1C. For the medium abundance clusters, those with 5+ spectra, the results, in Figure 3.1D and E, show that StrataCluster is able to maintain the low mixing and splitting, while also producing the highest number of distinct representatives when cluster representatives are searched in panel E. This shows that the stratified clustering process is both, sensitive and specific even when only considering the cores of clusters.

Cascade search

StrataCluster results thus reveal that as much as 94% of the human proteome is currently missed by typical data analysis protocols, including 69% of medium-abundance peptides, many of which occur in multiple tissues as seen in Figure 3.1B. To explore the space of possible identifications for precursors that are routinely not identified by typical search protocols (i.e, the MSGF+ protocol reported above) requires considering that peptides can be modified with atypical/unexpected post-translational modifications or can have slight sequence variations such as non-tryptic sequences or sequences with single amino acid polymorphisms. We explored this search space in two ways: (i) cascade search allowing for unexpected modifications and (ii) using spectral networks algorithms to define “families” of correlated peptide spectra (regardless of whether they’re identified or not).

To assess this missing proteome, it is important to first confirm what can be confidently identified in standard search protocols. Considering just unmodified peptides and peptides that are modified with known sample handling modifications, just over 1/3 of clusters are identified. Considering the spectrum level, propagating representative IDs over spectra, it appears that 50% of spectra in medium abundance clusters are identified. The discrepancy between spectrum counts and cluster counts here indicates, as expected that peptides commonly seen in other experiments make up a significant portion of spectra considered but add less to the overall count of peptides, showing the importance of clustering in assessing and prioritizing what is missing. The rest of the identifications are in the “gray proteome“, unexpected modifications and the “dark proteome“ or cases that are completely unmatched to anything found in the human proteome.

Unidentified proteome

The “gray proteome“ then consists of about 31% of medium abundance clusters, with 16% or 246,785 being identified by an open modification search. Many modified precursors indicate a broad diversity of precursors missed in typical experiments. Further, nearly all the additional open search clusters have modifications found are not those in the search space of MSPLIT/MSGF+ of common sample handling mods, such as methanine oxidation and N-term acetylation, and instead MODa is finding the majority mods that are post-translational, chemical derivatives, and amino acid substitutions, accounting for more than half of all modified clusters.

The second part of the “gray proteome“ wasn’t identified by open modification search but are cases where there are clusters that are in peptide families with identified spectra. While unidentified, nearly 60% of the time have a direct neighbor that is identified indicating that the aligned spectral similarity is likely sufficient to understand what the potential identification might be, with just one unexpected modification beyond what was in the original search space. Since it is part of known peptide families, these clusters are considered unidentified diversity and can help us understand the kinds of modifications that occur on these peptides (based on the delta mass of the edges). The remaining 40% of clusters have an edge to a cluster that is not identified, either due to other modifications – likely indicating something that is highly modified and unlikely to be found without spectral networks analysis, but still traceable to something identified.

24% of the remaining unidentified clusters are still in peptide families, just without any identifications in the network, or unidentified networked, making up part of the “dark proteome“. For these clusters, the topology of the network reveals “peptide families” even when none are identified. The likely scenarios include clusters are cases of sequences and modifications either from missing sequence space, or potentially spectra which when modified have substantially different fragmentation patterns than their unmodified counterparts.

Understanding tissue specific expression

Using information from spectral networks, identifications for unidentified clusters that were in the “grey proteome“ can be revealed, allowing for both an accurate assessment for the total abundance of a protein as well as an understanding of modifications, when there are potentially more than would be reasonable to find in a standard open search. Figure 3.3 shows the diversity of putative modified versions of sequence DQNGTWEMESNENFEGYMK which is uniquely mapped to protein P50120 and its isoforms, gene RBP2. RBP2 is a retinol-binding protein with a fatty acid binding pocket and is implicated in a host of roles in metabolism[62]. DQNGTWEMESNENFEGYMK is a peptide at the start of the protein which covers part of the binding pocket. The network of diversity around this peptide originally has only a few of the clusters identified, leaving many large clusters, containing 19 identified clusters out of 47, corresponding to an 1174 identified spectra, out of the total 2714, leaving the majority of spectra in the network unidentified, with likely most spectra from this peptide having 2+ modifications.

The identified spectra, those being from unmodified peptides and peptides with a smaller number of modifications occur mostly in the illium, jejunum, and duodenum (illium and jejunum are both considered small intestine in the dataset metadata). However, many of the unidentified spectra with propogated IDs over the network appear tissue specific to the small intestine, and appear to have labile mods. While follow up analysis is necessary to understand exactly what is happening, the added labile mods could potentially correspond to bound ligands at the binding pocket that are more prevalent deeper into the small intestine as compared to at the boundary with the stomach.

Exploring tissue-specific expression for all clusters

To begin to understand the “dark proteome“ , differential sample-level expression can provide clues as to what an unidentified cluster might be and help to prioritize future directions. At the protein level, the initial analysis of this draft proteome showed that around 50% of proteins appear to be housekeeping, or expressed in almost all tissues, many are tissue or group-specific1. At the peptidoform level, we can consider both unmodified expression, likely corresponding to protein-level expression, as well as expression of unexpected modifications, cases where some specific proteoform might only be present in a select number of tissues, even if the most abundant proteoform is not.

We consider “hyper-expression” of peptides, defined here as having over 10x the spectral counts for one tissue as compared to the median spectral count for a cluster. The cases of hyperexpression for unmodified spectra mostly fall in known proteins that have previously been shown to be tissue specific such as many peptides from sp|P12883|MYH7_HUMAN in heart tissue and sp|P01266|THYG_HUMAN in thyroid tissue.

Nearly half of the 81,791 of hyper-expressed clusters are unidentified, corresponding to highly abundant and tissue specific, and the relative identification rate of the tissue-specific hyperexpression varies significantly from tissues like heart and brain, where nearly all the hyperexpression is identified at the spectrum level (Figure 3.5B) to cases like bone marrow and lung, with most of the hyperexpressed clusters being unidentified. Further, we see that number of spectra continues to still not imply molecular diversity in this set, as the percentage of unidentified hyperexpressed clusters and unidentified corresponding spectra can differ significantly, both within the same tissue and between tissues, as stomach has more spectra in hyperexpressed clusters than thyroid, but thyroid has twice as many clusters as stomach. This discrepancy shows that StrataCluster is able to assist in assessment of follow up for all types of unidentified spectra, from cases in networks with identified spectra to networks with unidentified spectra.

Similar hyper-expression can be found in modified clusters, indicating tissue specificity for specific proteoforms as in Figure 3.5B. The peptide YI(C,305.076)ENQDSISSK from the protein sp|P02768|ALBU_HUMAN, a protein which is highly expressed in liver tissue [63] but also expressed in all tissues, is hyper-expressed in both lung and stomach tissue shown in Figure 3.4B. The modification is known to be an indicator of oxidative stress as it forms disulfide bonds with itself and with proteins to release two hydrogens that bind free oxygen to form water, thus removing “free-floating” oxygen thus reducing oxidative stress[64]. Additionally, this modified cystine is an important binding site on albumin, one of 4 disulfide bonds for Albumin binding site I, binding to drugs such as aspirin, bilirubin, azapropazone, warfarin, phenylbutazone, and tolbutamide[65].

Discussion

Lower bound clustering is a necessary tool in the systematic effort to understand the human proteome. To further what is yet to be understood, it is imperative to have an accurate count of the unknown. Further, breaking down the unknown by tissue we can then focus effort on identifying and understanding issues from tissues. From these unidentified peptides in common tissues there are a multitude of follow up experiments that can be run to validate these, from returning the set as a spectral library for further experimentation and searches to de-novo assembly. And by clustering in a low-redundancy manner, it allows for better spectral networks which can lead to better de-novo interpretation of the spectra[59].

Additionally, while we are currently using a lower bound clustering method and not considering the mixture strata in each cluster, we could develop a method that could deconvolute mixtures in clusters and cluster those again to reclaim more identifications. Through the deconvolution of mixtures, we could gain more identifications than searching the mixtures on their own as well as gain understanding of which peptides commonly coelute to guide experimental design to avoid this.

Another direction is towards complete repository clustering. Instead of looking at clustering homogenous experiments of 80 million spectra, we consider the entirety of MassIVE, combining thousands of experiments with a total of billions of spectra that are more heterogeneous, including spectra from many instruments, fragmentation types, and collision energies likely requiring new ideas in the similarity metrics but should be able to maintain the same framework for clustering. Finally, treating this problem as completely unsupervised is not necessary as there are now statistically controlled identifications at the repository scale[4] which would allow us to start the clustering knowing some IDs and using that to help understand what is left unidentified at the repository scale.

Supplemental methods

Human proteome data

To assess the clustering algorithm and compare to previously published tools, the analysis is done using spectra from a public dataset deep LC-MS analysis of 30 histologically normal human tissues, PXD010154 / MSV000083508 [26]. This dataset consists of 80,831,472 MS/MS spectra over 1,934 files in 2.12TB, but our analysis was restricted only to spectra from trypsin-digested proteins and fragmented using higher energy collisional dissociation (HCD), resulting in 68,140,357 MS/MS spectra used for all clustering experiments. The dataset was a partial submission, and not submitted with identifications.

Mass spectrometry selection of precursors for tandem (MS/MS) mass spectrometry

Tandem mass spectrometry is used to provide additional evidence for the existence of a molecule for precursor ions by further fragmentation. However, due to the presence of C13, N15, and other naturally occurring isotopes, the signal for a given charged peptide (precursor) does not exist completely in a single ion. For this reason, the mass spectrometer collects not just a single peak, but a small window around the selected peak, to capture all isotopes and increase the signal in the MS2. The set of all isotopes as well as the monoisotopic mass is called an isotopic envelope. The theoretical isotopic envelope can be modeled using the presence of C13 in averagine, an average amino acid composition[66]. Each peak will be 1/z apart, as that is the m/z of the difference of a charged C13-C12. The observed isotopic envelope in the MS1 will differ slightly due to the chemical composition of the peptide, as it differs from averagine, and also due to measurement errors.

The single selected precursor can be joined by other precursor ions that elute at the same retention time and with m/z within the captured isolation window, called coelution. Coelutions occur differently depending on chromatography and can vary across experiments, and samples that are more similar such as the same tissue will have more similar coelutions as compared to samples that are more different, such as from different tissues or other experimental differences[67]. The coelution causes an MS/MS spectrum to not contain fragments from a single precursor, but potentially from a set of precursors. When an MS/MS spectrum contains multiple precursors, this is defined as a mixture spectrum.

Mixture spectra are highly prevalent in the spectra from MSV000083508. Considering the explained intensity in the isolation window (IWEI), defined to be the sum of the intensity of peaks from the selected precursor ion and the other isotopes divided by the total intensity of all ions in the isolation window (in the case of this dataset, +/- 0.85 m/z around the selected ion), Figure 3.6 shows that more than half of identified spectra (see Clustering Evaluation section). have an IWEI < 0.8. This indicates that a large portion of the dataset has mixtures.

Mixture spectra cause difficulty for clustering as there is no way to directly assess the precursor ion for each fragment ion. Therefore, when calculating a spectrum-spectrum similarity, the standard assumption that this translates to a precursor-precursor similarity cannot always be assumed.

Kullback-Leibler (KL) divergence

The Kullback-Leibler (KL) divergence function can be used to hypothesize if an isolation window contains ions for more than one precursor without needing identifications. To use the KL divergence function, both the observed isotopic envelope and theoretical isotopic envelope need to be transformed into probability distributions.. To transform the theoretical isotopic envelope into a probability distribution, Q, the L1-norm of the envelope is taken and then each normalized peak intensity is added to the distribution, with an additional zero intensity before the first peak, between each peak, and after the final peak, as in Figure 3.7-1A. The observed isotopic envelope is transformed to the distribution, P, in a similar fashion. The difference is that all peaks outside beyond the locations for where the theoretical peaks fall are added to the distribution before, between, and after, as in Figure 3.7-1B.

In the calculation, peaks outside the theoretical isotopic envelope are penalized both in having intensity in an unexpected part of the distribution and using intensity that would otherwise have been from parts of the distribution where they should be matched. A perfect score, indicating that there is a single precursor in the isolation window is 0.

To assess the meaning of the KL divergence, only identified MS2 spectra (see Clustering Evaluation section) are considered. For identified MS2 spectra, the amount of mixing can be approximated by the explained intensity of the given peptide, the intensity of peaks that can be matched to fragment ions from the identified peptide over the sum of all peak intensities in the MS2. Using this identification information, Figure 3.8-A shows a violin plot of the empirically-estimated MS2 explained intensity density for the binned KL divergence of each spectrum. While there is a trend showing that spectra with higher KL tend to have lower MS2 explained intensity, this can be use to estimate mixing, but not an exact measure.

Further, as in Figure 3.8-B, even for cases where the average MS2 explained intensity is 40%, KL 2, 14% of precursors are lost when considering only spectra with KL < 2. While most of these spectra likely are mixtures, not all are, so a KL filter alone is likely too conservative to use prior to clustering, but can be useful as an additional consideration.

Another application of KL divergence can be to choose the correct monoisotopic peak for a spectrum. In the dataset MSV000083508, 2.9% of the time the monoisotopic mass chosen by the instrument did not match the monoisotopic match of the ID. To use KL divergence to determine the peak, the empirical distribution P needs to be slightly altered to not consider intensity between isotopic peaks, and then renormalized. If the envelope score is high, meaning the observed isotopic envelope does not resemble the theoretical isotopic envelope, either the precursor or charge was determined incorrectly and the different monoisotopic peaks can be considered. Using the highest scoring monoisotopic peak, 70% of these incorrectly chosen monoisotopic peaks can be corrected.

Dissimilarity edges

While KL divergence can help to identify mixture spectra, it is not specific enough to be useful in removing mixtures completely. As such, it is necessary to directly model the relationship between two spectra, with the expectation that there will be mixtures in the input set. StrataCluster introduces the idea of a dissimilarity edge, a directed edge which represent the containment of the some j of the top peaks of one spectrum k into all the peaks of the other. If the edge matches in only one direction, that is called an unidirectional dissimilarity edge and matches in both directions it is a bidirectional dissimilarity edge (see Figure 3.9).

MS2 spectra fragmented with HCD often contain non-discriminating peaks below 250 m/z, including immonium, a1-2, and b1-2 ions. Removing all peaks below 250 m/z before constructing dissimilarity edges ensures spectra that might share just these ions are not matched. For the analysis of MSV000083508, setting k to 5 and j to 4 leaves very few false edges while still maintaining high recall, for both unidirectional and bidirectional edges. Figure 3.10 shows this precision and recall for both A. bidirectional and B. bidirectional and unidirectional edges, when considering all pairs between spectra that share the same ID.

Constructing a graph where spectra are nodes and the edges are dissimilarity edges (within a parent mass tolerance) as defined above will greatly number of future similarity operations for each spectrum, without causing too many split clusters. Often times reducing the false positive rate by over 1000x for cases where many spectra have the exact same parent mass (see Figure 3.11.

Clustering the Dissimilarity Graph

To cluster the dissimilarity graph, StrataCluster considers each connected components created from the above dissimilarity graph separately. The following steps are performed within a single component.

Pseudo-clique clustering

Many spectra are nearly identical, and a key step to StrataCluster is to first group these spectra together. The spectra grouped will all have the same parent mass and also bidirectional similarity, as defined above. While bidirectional dissimilarity edges can be extremely precise at $< 0.1\%$ error, these edges are not sufficient to completely discriminate between two spectra at a large scale, as with hundreds of millions edges in a clustering graph, 0.1% could quickly create an abundance of mixing. To avoid this, the cosine similarity or the dot product of two normalized vectors is added as an added criterion to refine the similarity.

To perform the initial grouping, the edges are first refined as above. Next, the un-clustered spectra in each connected component are first ordered by their best outgoing edge, based on cosine similarity, with the intention of seeding the clusters with spectra that are similar to each other as possible. Calculating precision and recall for edges on nodes that have at least one bidirectional edge, consider edges with a cosine threshold of 0.8 to ensure that only a small fraction (<.05%) of edges drawn are incorrect - at the cost of recall being on average 60% per peptide for all edges see Figure 3.12. While the recall is low, this will be compensated for in a clustering step where 80% of edges must match between clusters when joining to the pseudo-clique, since the recall is 60% over all edges at cosine 0.8, it is 80% recall per node, allowing for about 20% true matches that do not have an edge to them.

The steps are in pseudo-code below:

```
1. Create a set of clusters, $C$, that is initially empty
2. For each spectrum, $s$, in all spectra in the connected component, $S$, with the same parent mass as $s$ that is not in a cluster, in this order do the following.
   (a) Form a new cluster, c, with spectrum, s, as a spectrum in the core of that cluster.
   (b) To add to the cluster. Let S' be all spectra that have a bidirectional dissimilarity edge to s.
       i. For all $s'$ in $S'$: If $s'$ has a bidirectonal edge to all spectra in $c$, and an edge with cosine greater than 0.8 to $\leq$ 80% of all spectra in $c$, add spectrum $s'$ to $c$ and remove $s'$ from the set of unclustered spectra.
   (c) Finally, to find the representative of c. Consider pairwise similarity between all spectra, s. The spectrum that has the highest average pairwise similarity is set to the first representative of c.
   (d) Update the core of c to the new s and add c to C.
```

Relaxed clique clustering

After the previous clique clustering step, all spectra which have nearly identical fragment peaks and have the same parent mass are now in the same cluster. However, due to different permutations in peaks and parent masses that can differ by an isotope, some spectra will not be clustered together. In the next phase, the previous stage is repeated, but using the representatives from before, and allowing for one isotope difference in parent mass between representatives to match.

1. Order all the clusters, C, by size.

2. For each cluster c in C, do the following.

    (a) Consider all clusters c' which are connected to c via the representative of c and the representative of c' by having a bidirectional dissimilarity edge (same process as above, just this time only with representative spectra) between the two representatives. If all the representatives of c' have a bidirectional edge to 80% of all representative spectra in the core of c, add cluster c' to c by adding the representatives of c' to c and add all spectra s' to the core of c. Remove c' from C. Repeat until all clusters c' are considered.

Affinity clustering

After the two rounds of clique clustering the unidirectional dissimilarity edges have not been considered yet, so many clusters that contain other clusters will be unconnected. To finally merge these clusters and complete the StrataCluster algorithm we apply an affinity clustering step. For unidirectional dissimilarity edges, projection cosine, an unidirectional measure of similarity[58], is used to refine the matches. Here, the measure is of the intensity of the spectrum containment, even if the full spectrum cosine is not high. The plots below, for both precision and recall for a projection cosine threshold illustrate the choice for these parameters, as at 0.6 projection cosine yields the maximum precision for the highest recall value (Figure 3.13).

Finally, in the StrataCluster algorithm, the assumption is made that larger, more homogeneous clusters are more likely to be the highest quality spectra for the particular peptide, and should be the cluster representative. Picking the cluster representative then determines if the edge created is a subset (the spectrum with more peaks is a mixture) or a superset (the spectrum with fewer peaks is missing signal). If there is a tie, the representative is chosen using minimum KL divergence for superset edges, and maximum precursor intensity for subset edges.

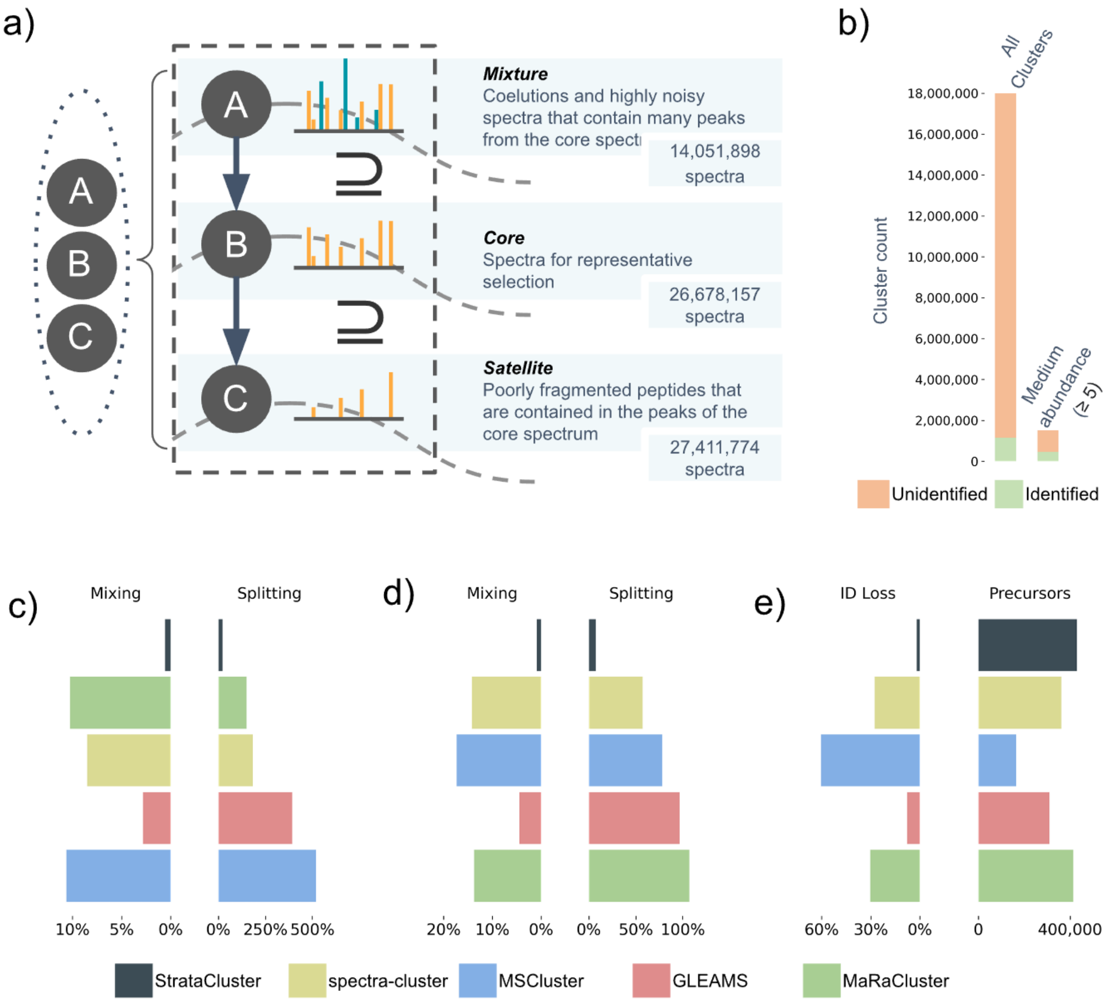
1. Order all the clusters, C, by size of the core.
2. For each cluster c in C, do the following, repeating until the following steps do not merge any two clusters.
   (a) Let C' be clusters with at least one representative spectrum with edges to c.
   (b) For each cluster c' in C', score c' as the average proj cosine for edges between c and c' via the representatives of c and the representative of c'.
   (c) Select the c' with the highest score and add to c' to c as follows:
      i. c' becomes a satellite of c if the edges between c and c' are bidirectional.
      ii. c' becomes a weak satellite of c if the edges between c and c' are directional with c' being a subset of c.
      iii. c' becomes a mixture of c if the edges between c and c' are directional with c' being a superset of c.
      iv. Remove c' from C (connectivity to the selected c' is no longer considered)

The clusters are now finalized, and the first representative for each cluster is the cluster representative for downstream analysis.

Acknowledgements

Chapter 3, in part is currently being prepared for submission for publication of the material. Pullman, Benjamin S.; Wertz, Julie; Carver, Jeremy; Bandeira, Nuno. The dissertation author was the primary investigator and author of this material.

Figure 3.1. a) Representation of the stratified cluster. The cluster is broken into three strata with each representing different components, the core, which are the highest quality most similar spectra, mixtures, which are cases that contain most of the peaks in the core but also extra peaks, and satellites, which are mostly contained in the peaks of core spectra. Spectra are evenly distributed between all three strata, with only 40% of spectra in the core strata. b) Number of clusters identified for both all clusters, and clusters with 5 or more spectra, noted as medium abundance clusters. For all clusters, the ID rate is only 6%, indicating the extent of what is unidentified over all spectra. Over medium abundance clusters, the ID rate is 30%, much more typical of standard proteomics experiments. c) The mixing and splitting for five clustering tools, using all clusters, including singletons, and cluster identifications based on majority identification in the cluster. d) The mixing and splitting as in c, except now considering only medium abundance clusters. e) The ID loss for medium abundance clusters, considering number of clusters identified by at least one component PSM as compared to the number of clusters identified when the representatives are searched and also the number of variant precursors (ignoring potential +1 errors), showing StrataCluster maintaining the highest number of identifications, with also the lowest splitting.


a)
A
B
C
A
B
C
⊇
⊇
Mixture
Coelutions and highly noisy spectra that contain many peaks from the core specti
14,051,898 spectra
Core
Spectra for representative selection
26,678,157 spectra
Satellite
Poorly fragmented peptides that are contained in the peaks of the core spectrum
27,411,774 spectra
b)
All Clusters
Medium abundance (≥ 5)
Cluster count
18,000,000
16,000,000
14,000,000
12,000,000
10,000,000
8,000,000
6,000,000
4,000,000
2,000,000
0
Unidentified
Identified
c)
Mixing
Splitting
10%
5%
0%
0%
250%
500%
d)
Mixing
Splitting
20%
10%
0%
0%
50%
100%
e)
ID Loss
Precursors
60%
30%
0%
0
400,000
StrataCluster
spectra-cluster
MSCluster
GLEAMS
MaRaCluster

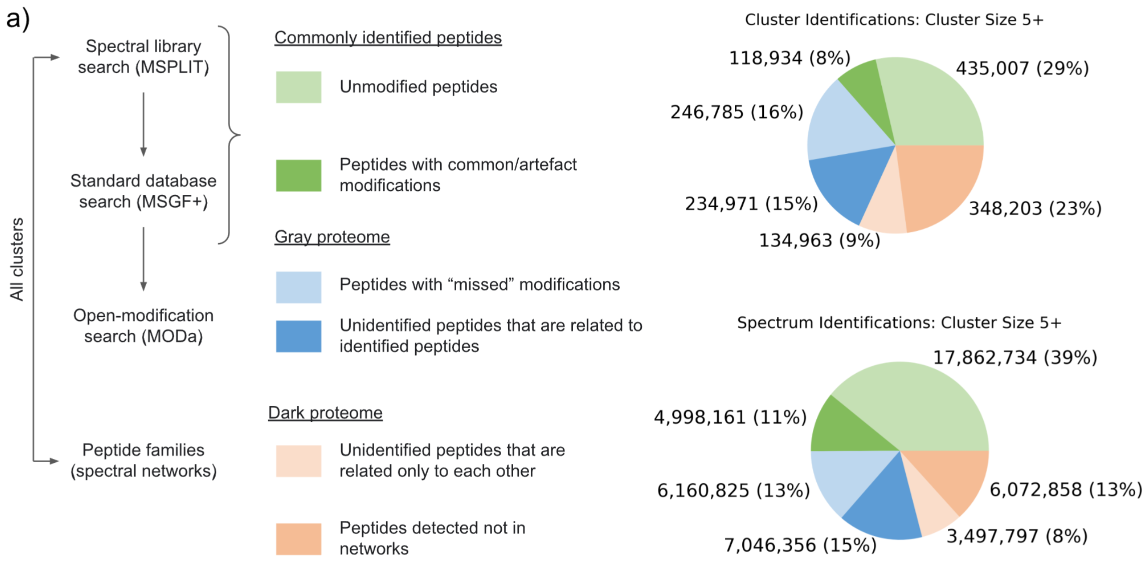


Figure 3.2. a) Maestro cascade search allows for an understanding of the medium abundance clusters, where first library search identifies clusters using MSPLIT searching the MassIVE-KB library. Clusters that were not identified by MSPLIT are then run through MSGF+ searched against the reference proteome. Finally, the yet unidentified clusters are run through an open modification search using MODa on a smaller database of proteins previously found in either MSGF+ or MSPLIT. Spectral networks are then constructed for the clusters. b)From the identifications and spectral networks, the clusters are filed into three categories, commonly identified peptides, which correspond to 37% of cluster identifications but nearly 50% of spectra from the clusters, then the gray proteome, which is peptides that are not in standard search spaces but which can likely be identified, and finally the dark proteome which is peptides where there is no hint as to what the ID might be, corresponding to nearly 32% of clusters, but only 21% of spectra.

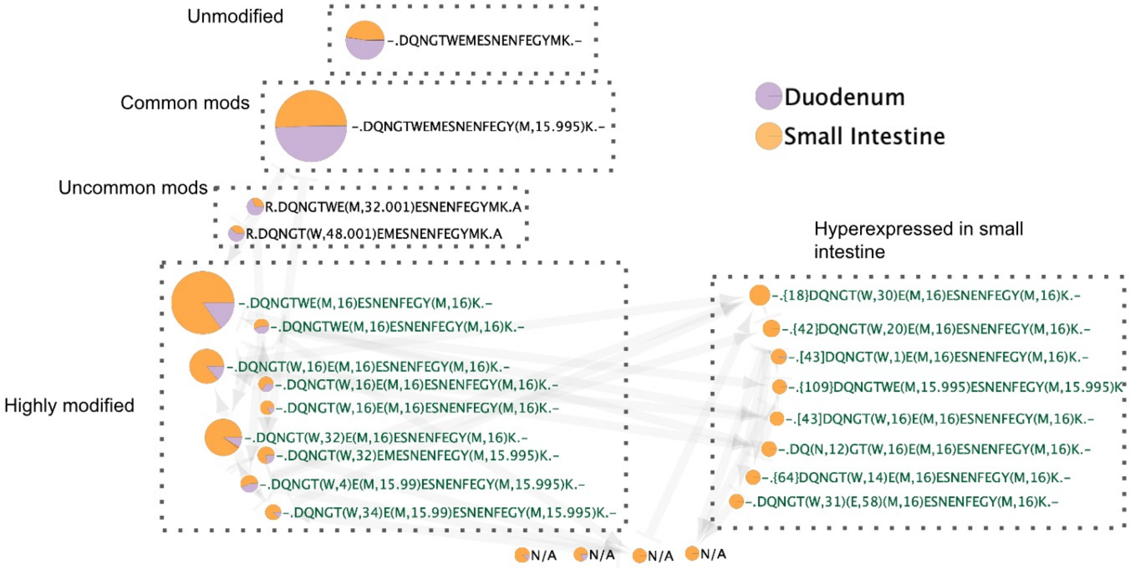

Unmodified
-.DQNGTWEMESNENFEGYMK.-
Common mods
-.DQNGTWEMESNENFEGY(M,15.995)K.-
Duodenum
Small Intestine
Uncommon mods
R.DQNGTWE(M,32.001)ESNENFEGYMK.A
R.DQNGT(W,48.001)EMESNENFEGYMK.A
Highly modified
-.DQNGTWE(M,16)ESNENFEGY(M,16)K.-
-.DQNGTWE(M,16)ESNENFEGY(M,16)K.-
-.DQNGT(W,16)E(M,16)ESNENFEGY(M,16)K.-
-.DQNGT(W,16)E(M,16)ESNENFEGY(M,16)K.-
-.DQNGT(W,16)E(M,16)ESNENFEGY(M,16)K.-
-.DQNGT(W,32)E(M,16)ESNENFEGY(M,16)K.-
-.DQNGT(W,32)EMESNENFEGY(M,15.995)K.-
-.DQNGT(W,4)E(M,15.99)ESNENFEGY(M,15.995)K.-
-.DQNGT(W,34)E(M,15.99)ESNENFEGY(M,15.995)K.-
Hyperexpressed in small intestine
-.{18}DQNGT(W,30)E(M,16)ESNENFEGY(M,16)K.-
-.{42}DQNGT(W,20)E(M,16)ESNENFEGY(M,16)K.-
-.[43]DQNGT(W,1)E(M,16)ESNENFEGY(M,16)K.-
-.{109}DQNGTWE(M,15.995)ESNENFEGY(M,15.995)K
-.[43]DQNGT(W,16)E(M,16)ESNENFEGY(M,16)K.-
-.DQ(N,12)GT(W,16)E(M,16)ESNENFEGY(M,16)K.-
-.{64}DQNGT(W,14)E(M,16)ESNENFEGY(M,16)K.-
-.DQNGT(W,31)(E,58)(M,16)ESNENFEGY(M,16)K.-
N/A
N/A
N/A
N/A

Figure 3.3. Example network where clusters with black text are identified in the cascade search and clusters with green text are identified by the propagation of identifications. Clusters with duplicate IDs because of charge, etc. are collapsed for easier readability. Relative abundance for each cluster based on spectra counts is represented by size of the cluster, and each cluster is split between either duodenum or other small intestine tissue. A lot of the spectral in the network comes from unmodified spectra and common modifications, but there is also a lot of spectra in clusters with 2+ mods, and additionally in highly modified clusters that are small intenstine specific.

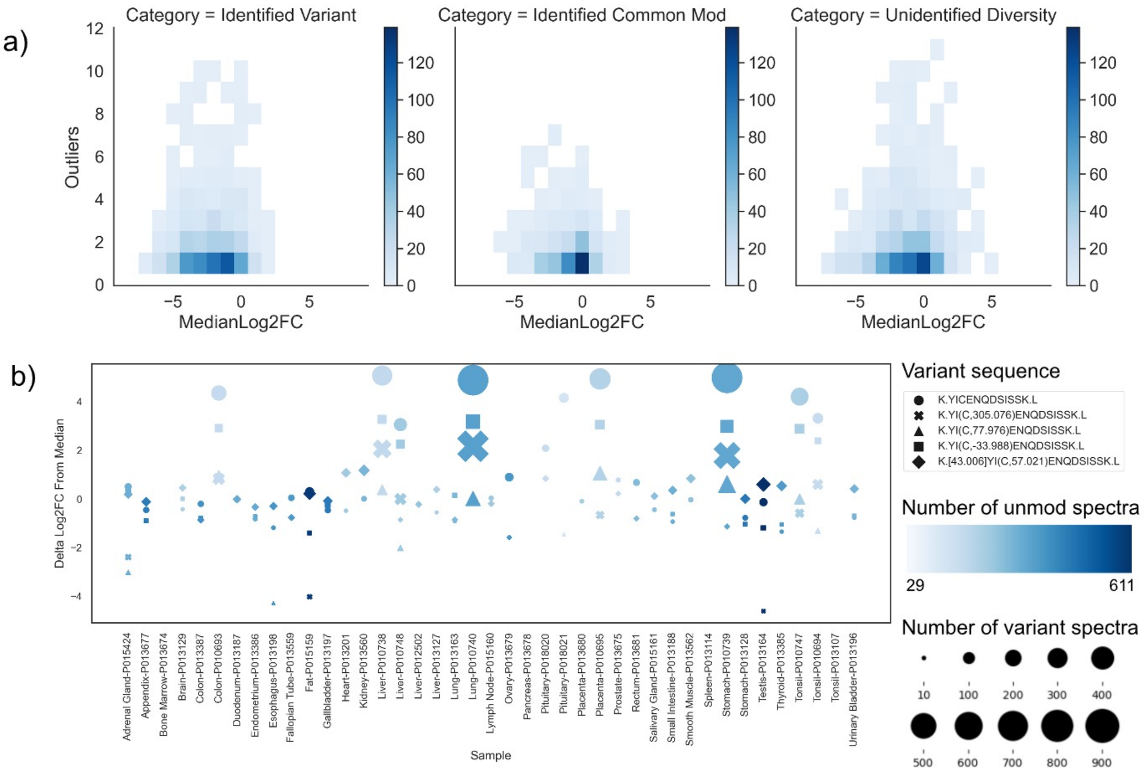


Figure 3.4. a) For each category of clusters, with a minimum of 10 samples per cluster and at least one outlier, defined spectral count expression 1.58 z-scores from the median, as the majority tend to have a median log2FC around 0, indicating not much change in expression between the unmodified sequence and the cluster variant. There are many outliers where the majority of log2FC is at 0, indicating that the expression is consistent in many tissues, however, there are a number of tissues where the expression is significant for that sample, as compared to the unmodified variant. B) An example of this is for the unmodified sequence YIC+57ENQDSISSK, 5 variants are considered and many variants include large clusters that have considerably different expression from the unmodified, here the variant is considered, and how much the expression differs from the unmodified in specific tissues, as compared to all tissues, taking into account both the number of unmodified spectra and the number of variant spectra. Some samples such as Lung-P010740 and Stomach-P010739 show significant expression changes

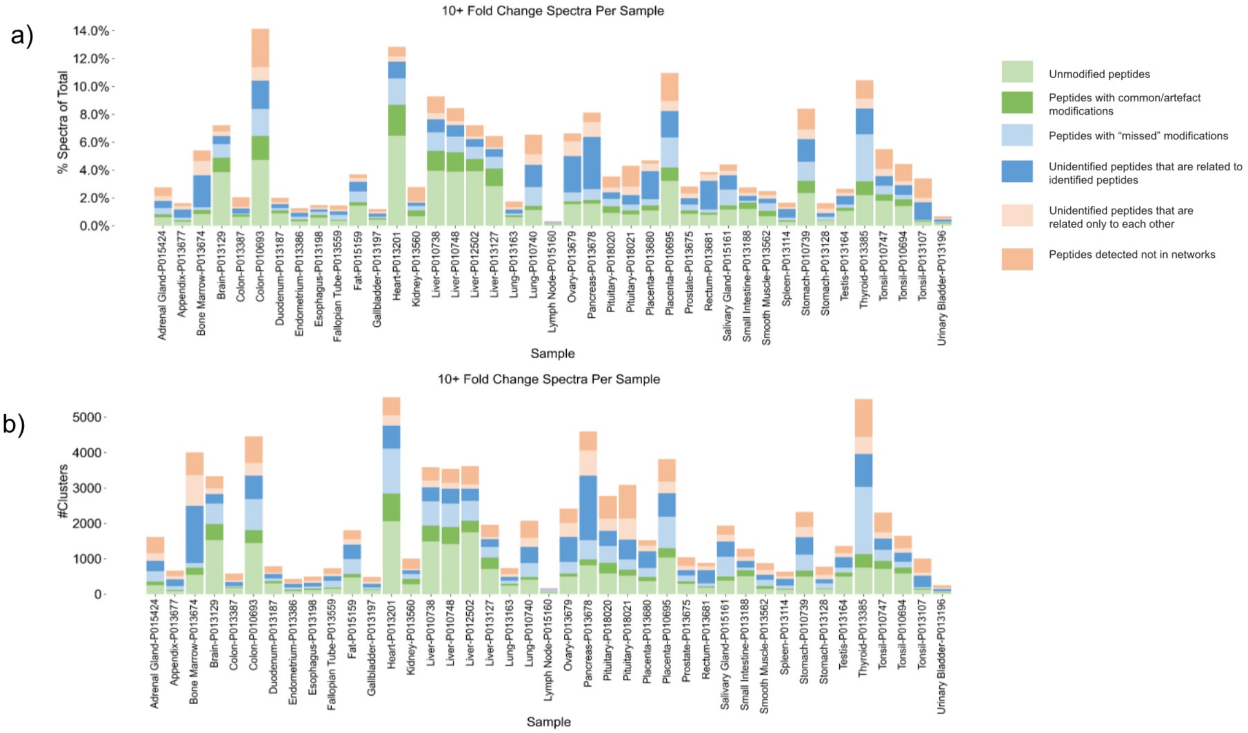


Figure 3.5. A. Percent of spectra per sample that are hyper-expressed (have a spectrum count for that tissue of 10x the median tissue count in the cluster). Each sample is broken down further by the percent of spectra that are in clusters in the known proteome, gray proteome, and dark proteome. B. Cluster counts for hyper-expressed clusters per tissue, these generally are similar to the spectrum counts.

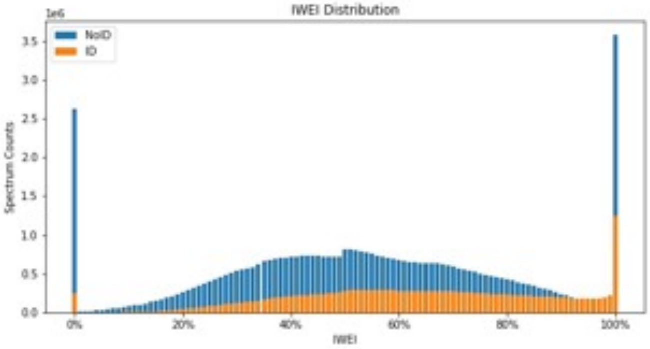


Figure 3.6. A histogram of explained intensity in the isolation window, using an isolation window of 0.85 m/z around the selected precursor ion, for both identified and unidentified MS2 spectra in MSV000083508.

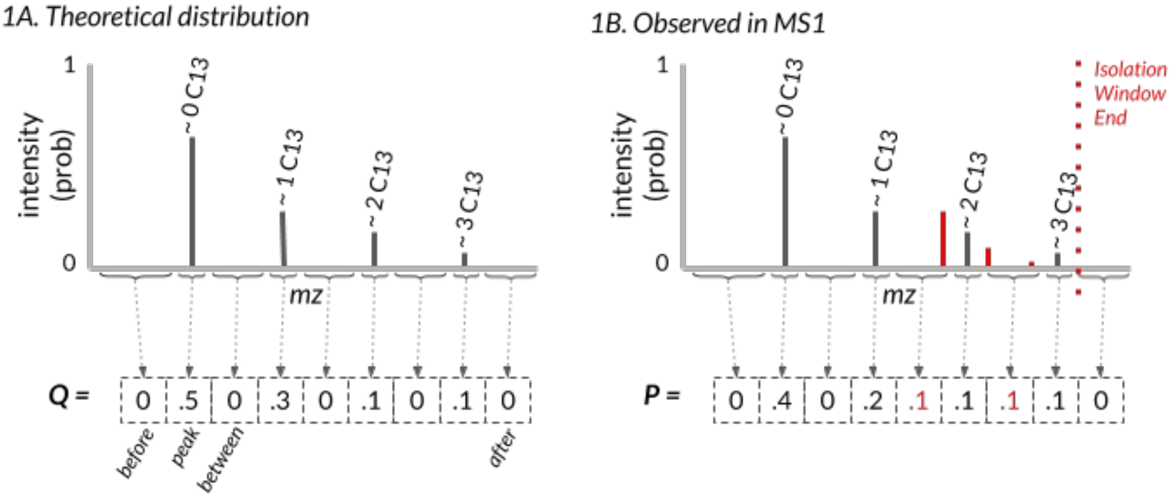


Figure 3.7. The left plot shows the theoretical distribution of isotopic peaks in an elution window, and right plot shows how coeluting peaks can occur

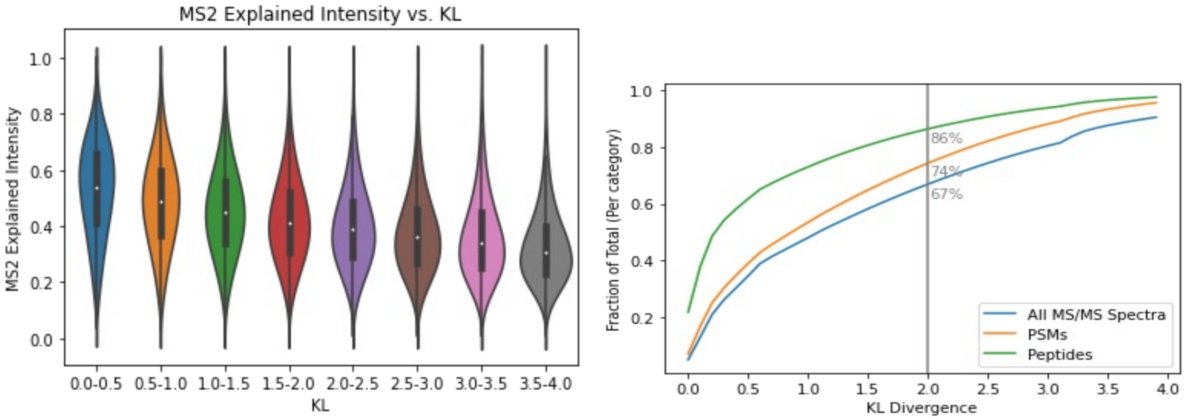


Figure 3.8. Left plot shows the MS2 explained intensity for different KL divergence values. Right plot shows the relative percentage of MS2 spectra, PSMs, and peptide for different KL divergence values

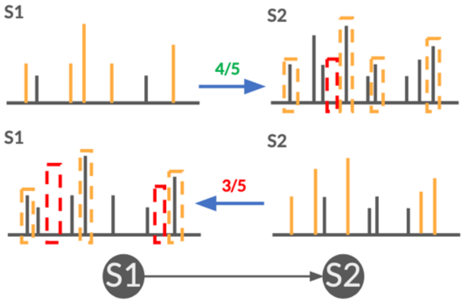


Figure 3.9. Illustration of a unidirectional dissimilarity edge with j=4, k=5. There are four peaks shared between the top peaks of S1 and S2, but only 3 of the top peaks of S2 are in S1.

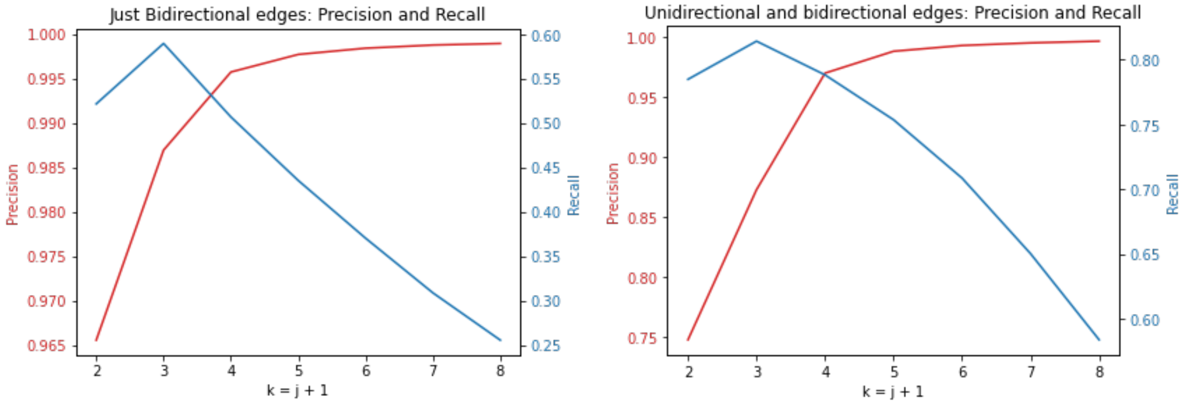


Figure 3.10. Precision and recall for considering only bidirectional dissimilarity edges and both unidirectional and bidirectional edges at different values of j and k ,where k = j+1. k=5 has the highest precision, without significant loss of recall.

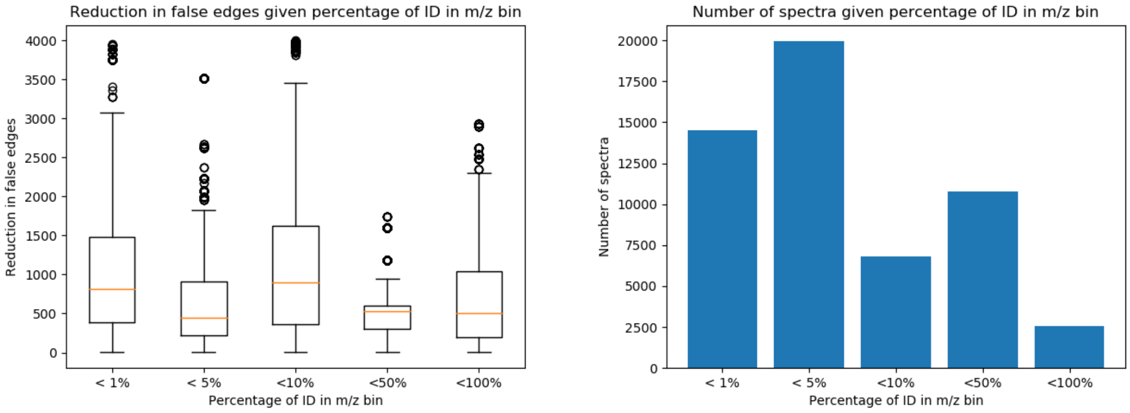


Figure 3.11. The left plot shows the amount of reduction in considered edges compared to all pairs, the right plot shows the number of spectra in each bucket. These are both taken from a subsample of the data.

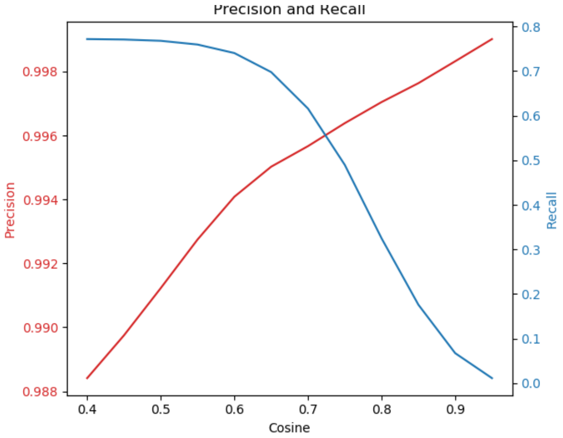


Figure 3.12. Precision and recall for different cosine thresholds for bidirectional edges. For the number of edges considered, the difference between 0.988 and 0.998 is significant. Note that all edges cannot be found determined with just bidirectional edges so the maximum recall is 0.8

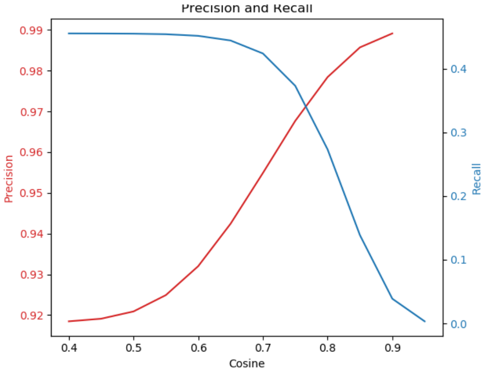


Figure 3.13. Precision and recall for unidirectional edges at various projection cosine thresholds

Chapter 4

Real-time modification-tolerant matching of MS/MS spectra at the repository scale

Introduction

How can the interpretation of newly-acquired tandem mass (MS/MS) spectra be informed by the billions of spectra acquired to date? This question is especially important for confirming the identification of surprising but important novel peptides/proteins, or for spectra that remain unidentified using standard methods. Furthermore, assessing the significance of novel identifications can benefit substantially from real time assessments of which tissues/datasets contain the same or modified/homolog versions of any peptide of interest. Conversely, repository-scale modification-tolerant matching is also an effective way to reject false positives by considering less-surprising interpretations of the same spectra as modified/homolog variants of otherwise commonly-detected peptides. We introduce a tool that enables these queries with near real time modification-tolerant searching against spectral libraries and public datasets.

While MASST[68] made it possible to query against hundreds of millions of spectra in the public GNPS[69] repository, this approach is not fast enough to be real time, requiring 2.5 CPU hours per open search query. The proposed approach is both able to query larger repositories, such as the 3 billion+ spectra in MassIVE, but also do so interactively, requiring usually less than 1 minute of user time for a standard open search query.

Methods

Spectrum notation

For each spectrum, S, define the spectrum parent mass as $S.pm$, and then for each peak $S[i] \in S$, define each fragment m/z as $S[i].mz$ and each fragment intensity as $S[i].int$.

Selecting files to run

The index is built with all publicly available datasets that are in MassIVE, including all public datasets submitted to GNPS. All open source spectrum peak files readable by ProteoWizard[70] in these public datasets are read into the index. It can be the case that two open source files are converted from the same RAW file, and care is taken to avoid using both by comparing file names and paths. When the index is updated, only files not considered in the original index will be used, as the workflow that finds which files allows for input files to be excluded.

Spectrum preprocessing and discretization

Before spectra are used in the indexing process, whether before they are added to the index to be searched against, or as queries, there is a preprocessing steps to improve the quality of the searches and to allow the spectra to inserted into or matched to the indexing data structures.

All preprocessing is applied to both the indexed and query spectra in an identical manner, unless the index spectra are known to be prefiltered, as is the case with spectrum libraries such as MassIVE-KB and the GNPS libraries, in which case no filters are applied but discretization steps are still applied. The first peaks to remove are the precursor peaks since they do not provide information about the spectrum fragmentation, where all peaks within 2 m/z of the precursor are removed and also peaks within 2 m/z of the neutral losses of H20 and NH3 are removed. Peaks below 200 m/z are removed in the proteomics indexes, to account for isobaric reporter ions, immonium ions, and other nondiscriminating ions, but no minimum mz peak filter is set for the small molecules indexes and this is configurable at the time of constructing the index. Next, a window filter is then applied, briefly, each peak p is only kept if it is in the top k peaks by intensity in the window from the $p.mz - r$ to $p.mz + r$, where r is the radius of the peak. In the indexes constructed, for proteomics data k is 8 and r is 50 and for small molecules k is 5 and r is 50.

Once the peaks are filtered, the spectra are discretized, by binning both the m/z for each peak as well as the intensity for each peak. This allows for both space saving in the index on disk as well as being a necessary step for the indexing algorithm. Peak mz values are discretized by first rounding each mz to the nearest multiple of 0.05 (this is a configurable parameter) for the construction of the index. If multiple spectrum peaks have their mz rounded to the same value then their summed intensities are assigned to the resulting peak with the rounded mz. To discretize the intensity, first a square root transformation is applied to all peak intensities to reduce the disproportionate influence of very high-intensity peaks (this is a configurable parameter) are then the peak intensities are normalized such that the spectrum vector's L2-norm becomes 1.

The dynamic programming (DP) recursions used for indexing compute the highest possible cosine to all theoretical spectra of Euclidian norm $1 = \sqrt{\sum_{s[i] \in s} (s[i].int)^2}$, which also implies that $\sum_{s[i] \in s} (s[i].int)^2 = 1$. As such, the DP recursions track the accumulated sum of squared peak intensities to guarantee that proper cosines are calculated between spectra of Euclidian norm 1. To achieve this, spectrum peak intensities are represented by their squared intensity when discretized into the bins used for peak indexing: a spectrum peak $s[i]$ with intensity $s[i].int$ is thus discretized into a bin $1 \leq n \leq p$ such that $\frac{n-1}{p} < s[i].int^2 \leq \frac{n}{p}$. The value p is an input parameter determining the resolution for representation of squared peak intensities and it can take any value between 1 and 255 (only one byte is used to represent each peak intensity). If the sum of bin-rounded squared peak intensities exceeds 1 for a spectrum then the spectrum is filtered to include only enough highest intensity peaks for the sum of squared intensities to total no less than 1.

Building the index

Once normalized, the spectra are grouped by parent mass. For each unit parent mass, all of the fragment peaks for each spectrum are combined and the peaks are bucketed first by unit fragment m/z, then by bucketed rounded fractional parent mass (at 0.05 m/z, there are 20 sub-unit buckets), rounded fractional fragment m/z, and index bucket, n, corresponding to $\lceil s.int \cdot p \rceil = \left\lceil \sqrt{\frac{n}{p}} \right\rceil$ in each indexed spectrum, s. From here, all spectra containing a parent mass, fragment m/z, intensity tuple are stored in an array, such that, $Index[n, mz, pm]$ points to the array of all spectra that have binned intensity of n, fragment m/z mz, and parent mass pm.

Querying the index: Finding the single peak bound

Once the peaks are normalized and the index is constructed, the next consideration is how to use the index to determine the set of candidate spectra that could possibly match with a cosine of at least θ to a query spectrum q with m peaks. As illustrated in Supp Fig DPTable, the DP recursions explore the space T of all possible theoretical spectra of norm 1 to determine whether a theoretical spectrum $t \in T$ exists such that $\text{cosine}(q,t) \geq \theta$. Initially T is unrestricted and thus contains all possible spectra, which guarantees that q itself is contained in T and thus a cosine of 1 is guaranteed to be possible (by definition of T).

The spectrum query algorithm is designed to iteratively constrain the space of possible theoretical spectra using an m-dimensional vector b, such that $t[i].int \leq b[i]$, for all $t[i] \in t$. Starting with $b[i] = 1$ for all $b[i] \in b$, the algorithm iteratively reduces the $b[i]$ bounds on maximum peak intensities to progressively define a restricted space T^R of theoretical spectra until two important conclusions can be established:

- No theoretical spectrum t exists in T^R with $\text{cosine}(q,t) \geq \theta$.
- All theoretical spectra t that could possibly have $\text{cosine}(q,t) \geq \theta$ must be in $T \setminus T^R$ and must thus violate at least one of the $b[i]$ constraints defining T^R (i.e., at least one peak $t[i] \in t$ must have intensity $t[i].int > b[i]$).

It is also important to note that since the cosine between q and t can only be affected by peaks at mz values with non-zero intensity in q, the algorithm only needs to consider theoretical spectra t with all of their intensity at the same mz values as q. If no such spectrum t exists in T^R that can have a cosine$(q,t) \geq \theta$ even when all intensity in t is in the same m dimensions as q, then all other spectra t^- such that $t^-[i] \leq b[i]$ and with norm < 1 in the same m dimensions as q will have even less intensity to match to the peaks in q and thus cannot possibly have cosine$(q,t^-) \geq \theta$.

The query algorithm is formally defined as follows:

- m: number of fragment peaks in q; $i = 1 \dots m$ for all equations below.
- p: number of bins used to represent squared intensities; $n = 0 \dots p$ for all equations below. All discretized squared intensities below are in the scale $n = 0 \dots p$ representing discretized squared intensity $0 \dots 1$. As such, for any peak $s[k]$ with intensity $s[k].int$ in any spectrum s, the corresponding discretized squared intensity is defined as $s[k].dint = \lceil p \times s[k].int^2 \rceil$
- b: a 1-dimensional array of size m, where $b[i]$ is the upper bound of squared peak intensity that theoretical restricted spectra $t \in T^R$ are allowed to use when matching a peak at $t[q[i].mz]$ to $q[i]$, such that $t[q[i].mz]].dint \leq b[i]$; b is initialized to $\forall_{i=1 \dots m} b[i] = p$ (representing the maximal squared peak intensity of 1). At the conclusion of the spectrum querying algorithm, any indexed spectrum s with at least one peak at m/z $q[i].mz$ and $s[q[i].mz].dint > b[i]$ is considered a possible match to q and is passed to the next stage for calculation of the full cosine.
- T^R: set of all possible theoretical spectra $t \in T^R$ with spectrum peaks $t[q[i].mz].int \leq b[i]$
- C: a 2-dimensional array where and $C[n,i] =$ number of spectra $s \in S$ with a peak at mz $q[i].mz$ and intensity $s[q[i].mz].dint = n$
- P: a 2-dimensional array of size $m \times p$ and

$P[n,i] = q[i] \cdot \sqrt{\frac{n}{p}}$

- L: a 2-dimensional array where $L[n,i]$ = maximum possible cosine between $q[1], \ldots, q[i]$ and all theoretical spectra $t \in T^R$ of sum squared norm n for all peaks matched to $q[1], \ldots, q[i]$

$$L[n,i] = \begin{cases} 0 & \text{if } n = 0 \\ P[n,1] & \text{if } i = 1 \\ \max_{n=k+j} L[k,i-1] + P[j,i] & \text{if } n > 0, i > 1 \text{ and } \forall_j 0 \leq j \leq b[i] \\ -\infty & \text{otherwise} \end{cases}$$

- R: a 2-dimensional array where $R[n,i]$ = maximum possible cosine between $q[i], \ldots, q[m]$ and all theoretical spectra $t \in T^R$ of sum squared norm n for all peaks matched to $q[i], \ldots, q[m]$

$$R[n,i] = \begin{cases} 0 & \text{if } n = 0 \\ P[n,m] & \text{if } i = m \\ \max_{n=k+j} L[k,i+1] + P[j,i] & \text{if } n > 0, i < m \text{ and } \forall_j 0 \leq j \leq b[i] \\ -\infty & \text{otherwise} \end{cases}$$

- M: a 2-dimensional array where $M[n,i]$ is the maximum possible cosine between q and all theoretical spectra of squared norm n with a peak of squared intensity i at m/z $q[i].mz$

$$M[n,i] = \begin{cases} 0 & \text{if } n = 0 \\ \text{max}_{1=k+n} R[k,i+1] + P[n,i] & \text{if } n > 0 \text{ and } i = 1 \\ \text{max}_{1=j+n} L[j,i-1] + P[n,i] & \text{if } n > 0 \text{ and } i = m \\ \text{max}_{1=j+k+n} L[j,i-1] + R[k,i+1] + P[n,i] & \text{if } n > 0 \text{ and } 1 < i < m \\ -\infty & \text{otherwise} \end{cases}$$

The process for finding the theta bound starts by constructing the cost matrix C. For each peak q in the query spectrum q (in Figure 4.3-1) with parent mass pm_q, consider all peaks in the index for all parent masses pm_i (such that pm_i is in the query range) that have fragment mass of n and also the shifted peaks $q + (pm_q - pm_i)$. Add the total count of spectra at each peak and intensity into the corresponding cell in C. When constructing C, there is an optimization to not use all parent mass bins. This is configurable but for the results in the indexes as constructed we use one Da bin per every 10 Da. There is a trade off between potentially missing a large bin as compared to the IO cost to load the index.

Next, each element in b is initialized to p (in Figure 4.3-2). From here, L and R are constructed, ensuring that the amount of intensity considered for each peak, i, is no more than the intensity in $b[i]$ (in Figure 4.3-3,4). While finding $L[n,i]$ and $R[n,i]$ involves a maximum over two variables, k and j, (and potentially another factor of p in the computation) this can be computed incrementally from sum of squared peak intensity bins 0 to n and peaks from 0 to i in L and m to i in R, when both k and j are saved for each bin. This optimization removes a multiplicative factor of p from the run time and works as the incremental contribution to the cosine only depends on the immediately preceding bins, as the rate of added contribution decreases monotonically as more intensity is contributed due to the binning of the intensities. After L and R, update M, using the updated intensities from L and R (in Figure 4.3-5). This can also be done incrementally so that the $best_{ln}$ and $best_{rn}$ do not need to be considered for each cell, as the added contribution from L and R also decreases monotonically similarity to L and R.

Once M is constructed, the next step is to determine which peak from the query to allocate more intensity (in Figure 4.3-6). For each peak, consider the cost in candidates for each decrement in cosine and pick the peak where the delta cosine divided by number of candidates is highest. Remove the number of now added candidates from $C[n,i]$, as to not count them twice. Repeat the process of constructing L and R and then M, until the maximum cosine in M is less than the user defined theta.

Querying the index: Using single peak bound to determine candidates

Once b is determined, the next step is to find all peaks from the index corresponding to fragments in the index and to construct partial spectra. For each query peak, i, there can be multiple matching m/z fragments in the index peaks at different parent masses (depending on input parameters). Set $P[i]$ to be the set of all fragments, parent mass tuples, (f, pm) in the index that have peaks that match $q[i]$ in the following two ways:

1. Fragment peaks that have an m/z that is within a predefined tolerance, t of the peak, such that $|q[i].mz - f.mz| \leq t$ regardless of parent mass differences.

2. Fragment peaks that have an m/z within tolerance and including potential parent mass offsets, where $d = q.pm - pm$ and $|q[i].mz - f.mz - d| \leq t$.

Construct an empty set of partial result spectra, R, where each spectrum r contains m peaks, initialized to all zero. These will be the partial candidates found from the index. From the query, for each peak, i from 0 to m, fetch all candidates from the peaks in the index as $Cand[i] = \forall_{(f,pm) \in P[i], n \leq b[i]} Index[n, f, pm]$. Within $Cand[i]$ it is possible for there to be cases where the same spectrum is included twice, as it is possible for multiple fragments to match because of a the peak tolerance or shifted peaks. In these cases, remove the redundancy by keeping only the peak which has the highest potential contribution to the cosine, that is with intensity closes to $q[i].int$ and disregard the other matched peaks. For each candidate in $Cand[i]$, add the matching peak intensity for the candidate at each $r[i] to R$, adding a new r to R if it was missing.

For each peak found, consider the matched intensity between the query and the candidate and calculate a upper bound on the potential cosine for the candidate r. There are three ways in which the cosine can be incremented for each peak i. This allows us to calculate partial cosines using the bounds from b, even if there is no matching peak at $r[i]$, as the most conservative estimate in all cases can be considered, if enough information is known about $b[i]$, the intensity can be estimated to be the highest amount of unseen intensity. Using this estimate of the cosine upper bound, any partial spectrum r having estimated cosine less than θ does not need to be considered further and will be removed from R. Note that sometimes it can be more efficient to use a $\theta - \varepsilon$ bound instead of θ, as the extra candidates included in the additional ε amount of potential cosine might be worth including as a way to reduce b further, so that the upper bound on cosine can be more aggressive.

Querying the index: Filtering potential candidates using full spectrum information

For all spectra, r in R, that pass the previous filtering step, calculate the full spectrum cosine between the discretized spectra for both the filtered result, r and query, q, allowing for shifted peaks if allowed, and output all spectra that have a cosine greater than θ. These full spectrum peaks are organized per spectrum, rather than per fragment peak, can be directly indexed knowing the candidate number from the previous index. Once spectra with a full spectrum cosine $> \theta$ are known, metadata about each spectra can be indexed by the candidate number similar to the spectrum peaks. For the massive index, metadata included is the modified peptidoform and also the charge in MassIVE Search, as well as MassIVE-curated tissue mappings per file.

Disk-based index data structures

All of the above processing can work with the data structures in memory, or also on disk. Each index consists of seven categories of files on disk. The first, referred to as the "peak index”, contains spectra organized based on the peaks in the spectrum, where each parent mass, fragment mass, intensity tuple is associated with all spectra containing that tuple. The second, referred to as the “spectrum index” contains the all the peaks from each spectrum, grouped by spectrum (and not by peak mz/intensity). The third, the “spectrum annotation index” contains metadata about each spectrum. The fourth, the “filenames index” contains a list of all peak list files used in the index. The fifth, the “annotations” is a non-redundant list of all metadata terms used in the index. The sixth, the “file-level annotation index” contains metadata for each file, that applies to all spectra in the file. The seventh, the “processing parameters” provides a list of the preprocessing parameters used in the index.

1. **Peak index: *.mxc** The peak index consists of metadata plus two sections organizing spectrum peaks. The the metadata, containing the number of parent mass sub-unit bins (i.e. number of discrete bins between two units), the number of fragment mass sub-unit bins, and the number of index bins, and then a 2000 dimensional vector of fragment unit bins containing offsets to the index for each unit fragment peak. The two sections are each repeated once per unit fragment bin and are as follows (shown in Figure 4.4):

   (a) The first is an array of the pairs of peak triples and offsets, where each peak triple is a 32 bit unsigned integer (bits 0-7 are the intensity index, 8-15 are the fractional parent mass, and 16-23 are the fractional fragment tolerance, and 24-31 are set to 0). The second array contains all the spectrum indices for spectra that contain a peak in the unit parent mass, unit fragment mass bin, sorted by the peak triple for each index.

      i. The difference of consecutive pointer values indicates the number of candidates at a particular tuple of (sub-unit fragment mass, sub-unit parent mass, intensity bucket) (i.e. number of bytes is 4 times this difference)

   (b) The offsets from the first array correspond to the first position in this second array where the spectra indices have that particular index triple.

This allows us to only store the index for triples which are in the data, costing at most 3*peaks for each additional spectrum (adding an index + offset for each peak), though likely much less, as the peak likely is not completely unseen.

2. **Spectrum Index: *.mxs** These files (one per parent mass unit Da) contain information for all the spectra that have a parent mass in the unit parent mass for the file. There are two arrays in this file:

   (a) The first contains index-offsets to the start of all spectra in the second array, and is guaranteed to have the same number of elements as there are spectra in the index at the given unit parent mass.

      i. The difference of consecutive pointer values indicates size of spectrum to read (i.e. number of bytes is 4 times this difference)

   (b) The second array contains all the information for each spectrum as follows:

      i. Exact (non-binned, non-discretized) parent mass

      ii. Binned parent mass

      iii. Charge

      iv. Scan number

      v. File index number (referencing a line in filenames.txt)

      vi. Peaks: an array of 32-bit integers, where bits 0-7 are the peak intensity, and bits 8-31 are the binned and discretized fragment mz

3. **Spectrum Annotation Index: *.mxa** These files (one per parent mass unit Da) contain information about any annotations for the spectra. Generally, these will contain PSM information. The design mirrors the *.mxs files, with two arrays as follows:

   (a) The first contains offsets to the start of all spectra in the second, and is guaranteed to have the same number of elements as there are spectra in the index at the given unit parent mass.

    (b) The second array contains all the metadata for each spectrum, empty if the spectrum does not contain any metadata. However, if the spectrum does contain metadata, the metadata is represented as an array of triples containing of

        i. the offset in bytes in the annotations.txt file

        ii. the length in bytes of the annotation to read

        iii. the offset in line numbers in the annotations.txt file

   By representing annotations in a separate file from spectra we allow for changing annotations without needing to rewrite the large spectrum peak files.

4. **Filenames index: filenames.txt** One filename per line of all peak list files in the index. Filename is relative to where the code that constructs the index is run.

5. **Annotations: annotations.txt** One annotation string per line, annotation string can be anything.

6. **File-level annotation index: filenames.mxa** Similar to annotation files except providing metadata information per file, for cases where the file all has the same annotation and keeping track per spectrum could be inefficient.

7. **Processing parameters: processing.json** Processing parameters for the index to ensure queries are filtered as the index (can be overwritten with command line parameters):

    (a) Index version

    (b) Window filter parameters

    (c) SNR parameters

    (d) Minimum fragment m/z

Results

Timings

To time the queries we launched single spectrum search requests to the indexing search API by USI[47]. No spectra or index is preloaded or cached. The analog searches allow for matches from 130 Da below the query parent mass to 200 Da above. For the MassIVE searches, USIs from identified spectra in cHPP datasets are used and for the GNPS searches, GNPS library spectra are used. While all searches are done in real time, it is important to consider the type of searches and the reasoning behind which is used. As in Table 4.1, queries to libraries are all less than a second, and queries against repositories are less than a second for exact searches, and less than a minute for open searches.

Use case: Diversity of Detection

When presenting evidence for a novel protein existence, HPP criteria[36] require two, high quality, non-nested peptides for as evidence for the protein, as discussed in Chapters 1 and 2 of this document. While sufficient FDR controls and manual analysis can help confirm that the peptides are high quality, this repository scale search allows for a quick check among all repository spectra if a) this spectrum matched to the peptide is present in other datasets and if the fragmentation matches other datasets, adding weight that the peptide is likely real and reproducible and b) if a very similar spectrum is found in another dataset with a different ID, it calls into question the quality of the potential ID. For the protein sp|Q86X67|NUD13_HUMAN, a newly PE1 protein, we confirmed by searching against the repository for one of the peptides used to provide evidence for the missing protein, DASLLSTAQALLR. By searching one PSM for DASLLSTAQALLR, mzspec:PXD003947:QEx2_006887:scan:28591:DASLLSTAQALLR/2[11], we found over 80 matches with a cosine of > 0.7 in the MassIVE repository, most of which have highly correlated fragment peaks and 6 PSMs matched that are identified (at the time of publication many spectra in MassIVE are not yet identified as either they were not a complete submission or have yet to be reanalyzed, but with a recent version of MassIVE this changes signficantly) have the same sequence. However, two PSMs match with a very similar cosine, but to a different ID of GLDTISVTGNILR, and if we compare the mirror plot in Figure 4.1 we see that while many fragments are shared such as *y*7, *y*8, *y*9, there are enough high intensity fragments to justify that the original PSM is correct. While the PSM in question here was visually of good quality, many remaining MP will not necessarily be, so having the ability to search for consistent fragmentation could help in giving confidence to an ID for a difficult to call protein.

Use case: Finding Diversity and Occurrences of unidentified peptides

The indexing can also be used to help illuminate the dark proteome, showing which datasets unknown spectra appear in, as well as with what mass offsets. In the case of the spectrum, mzspec:PXD010154:01280_A05_P013164_S00_N33_R2:scan:2375[26], this is unidentified in the original analysis (Chapter 3), but has over 2000 matches to unidentified spectra in 40 datasets. Additionally, this spectrum matches 570 spectra with a delta mass of +78 in nearly all of the matching datasets, but some modification masses such at +46 are more dataset specific as in Figure 4.2. This repository-scale search provides a way to begin to prioritize cases from the dark proteome to examine.

Discussion

Efficient indexing and algorithms enable real-time, modification-tolerant, repository-scale searches against billions of spectra enabling the use of full repositories to help confirm or reject novel identifications and understand diversity of spectra. This search, as shown, can be used in the context of all chapters of this thesis, making it a useful tool in the future exploration of mass spectrometry data, summarizing the wealth of years of compute and thousands of researchers globally in real time analysis.

Acknowledgements

Chapter 4, in part, is a reprint of the material as it appears in "Real-time modification-tolerant matching of MS/MS spectra at the repository scale". ASMS 2021 - MP159. Pullman,

Benjamin S.; Batsoyol, Narangerelt; Swanson, Steven; Wang, Mingxun; Bandeira, Nuno and, in part, is currently being prepared for submission for publication of the material. Pullman, Benjamin S.; Batsoyol, Narangerelt; Swanson, Steven; Wang, Mingxun; Bandeira, Nuno. The dissertation author was the primary investigator and author of this material.

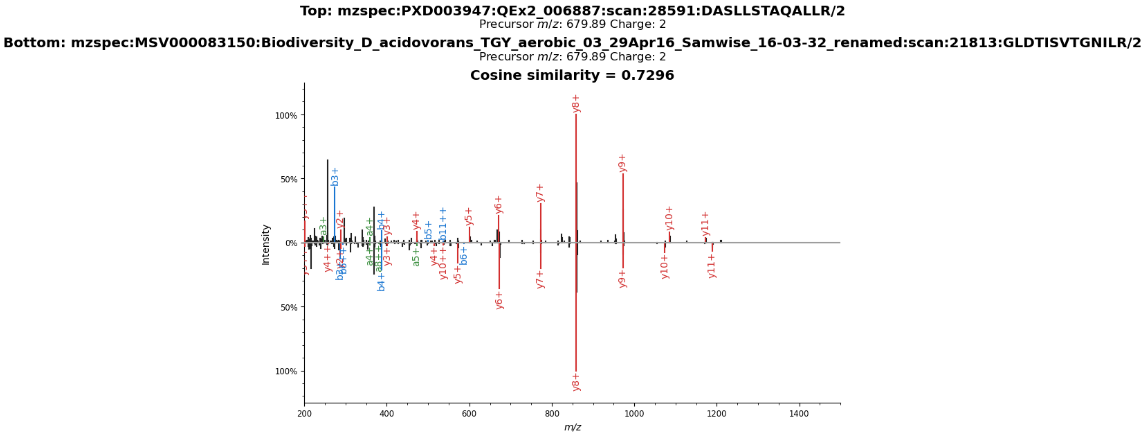


Figure 4.1. A plot[71] showing the a PSM for DASLLSTAQALLR and a PSM for GLDTISVTGNILR, which have a high cosine but enough unique peaks to likely be from different peptides and show evidence for sp|Q86X67|NUD13_HUMAN.

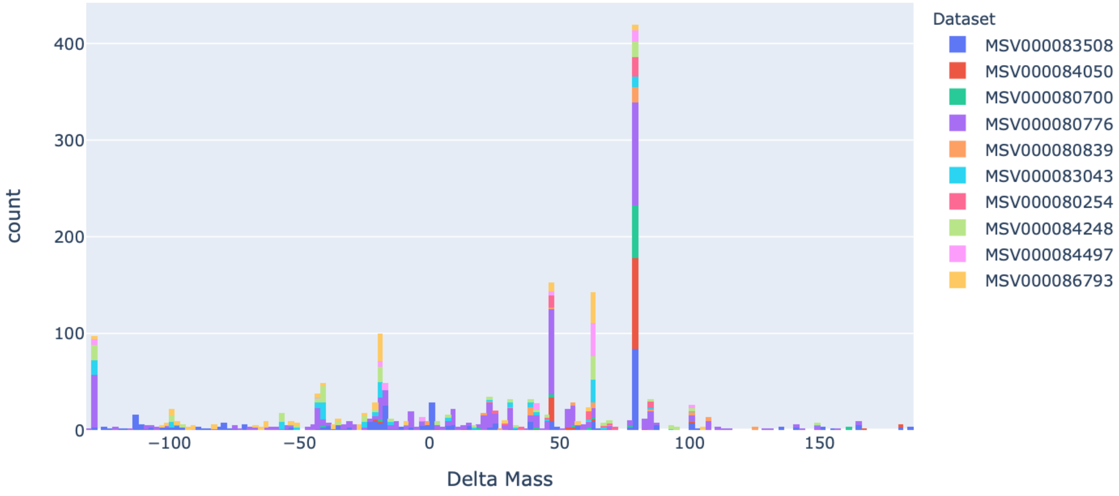


Figure 4.2. A histogram of all matching spectra to the unidentified spectrum mzspec:PXD010154:01280_A05_P013164_S00_N33_R2:scan:2375[26] showing the dataset and delta mass for each match.

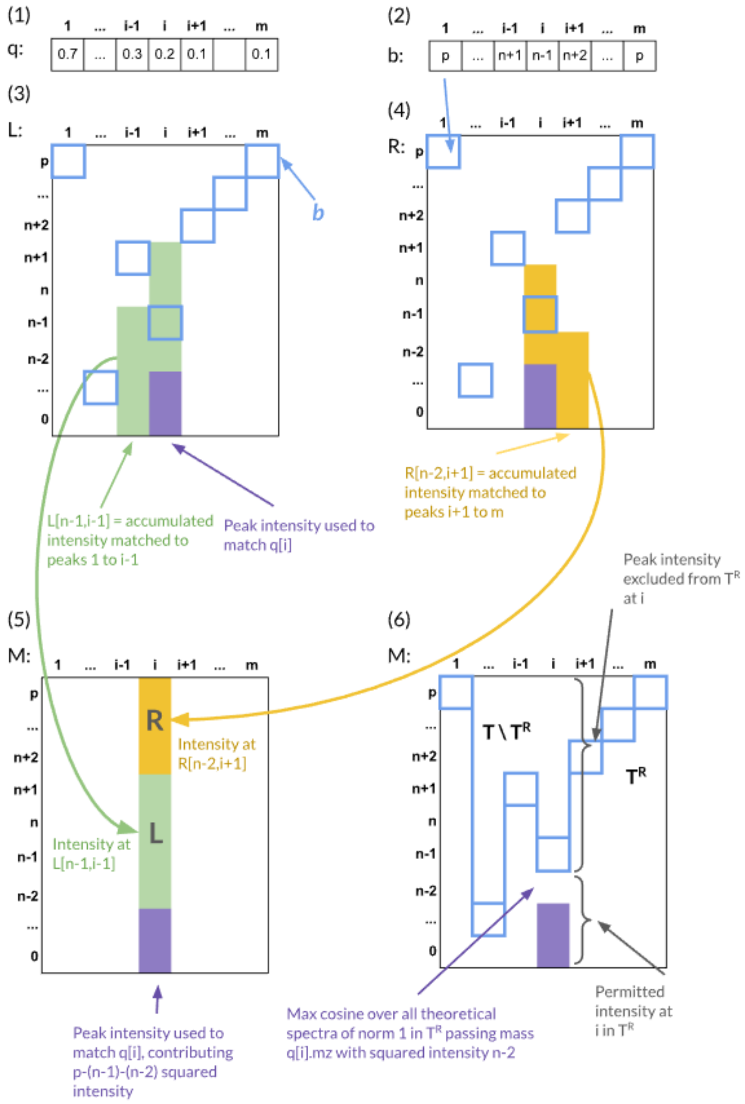

(1)
q:
1 ... i-1 i i+1 ... m
0.7 ... 0.3 0.2 0.1 0.1
(2)
b:
1 ... i-1 i i+1 ... m
p ... n+1 n-1 n+2 ... p
(3)
L:
1 ... i-1 i i+1 ... m
p
...
n+2
n+1
n
n-1
n-2
...
0
b
L[n-1,i-1] = accumulated intensity matched to peaks 1 to i-1
Peak intensity used to match q[i]
(4)
R:
1 ... i-1 i i+1 ... m
p
...
n+2
n+1
n
n-1
n-2
...
0
R[n-2,i+1] = accumulated intensity matched to peaks i+1 to m
Peak intensity excluded from T^R at i
(5)
M:
1 ... i-1 i i+1 ... m
p
...
n+2
n+1
n
n-1
n-2
...
0
R
Intensity at R[n-2,i+1]
L
Intensity at L[n-1,i-1]
Peak intensity used to match q[i], contributing p-(n-1)-(n-2) squared intensity
(6)
M:
1 ... i-1 i i+1 ... m
p
...
n+2
n+1
n
n-1
n-2
...
0
T \ T^R
T^R
Max cosine over all theoretical spectra of norm 1 in T^R passing mass q[i].mz with squared intensity n-2
Permitted intensity at i in T^R

Figure 4.3. The query vector q with m dimensions. 2. The vector b restricts the maximum intensity considered from index spectra that have a peak at i. 3. The matrix L defines the maximum accumulated cosine considering peaks from left to right. Here $L[n+1,i]$ uses $n-1$ intensity from peaks $1\ldots i-1$ and the remaining intensity from the purple peaks in the index matching query peak $q[i]$ at the given intensity. 4. The matrix R is the same as L except going from $m\ldots i$ and reporting the accumulated cosine from right to left. 5. The matrix M shows the maximum possible cosine for theoretical spectra of $L2-$norm 1. To calculate the maximum possible cosine for spectra passing through the purple peaks in the index at i, the accumulated cosine contributed from the peaks $0..i-1$ from the left and the accumulated cosine contributed from $i+1\ldots m$ on the right. 6. Once M is calculated, it is possible to know the maximum possible cosine from all spectra in T^R and once this cosine is $< \theta$, only spectra in the index containing a peak i, above $b[i]$ will be considered for further processing.

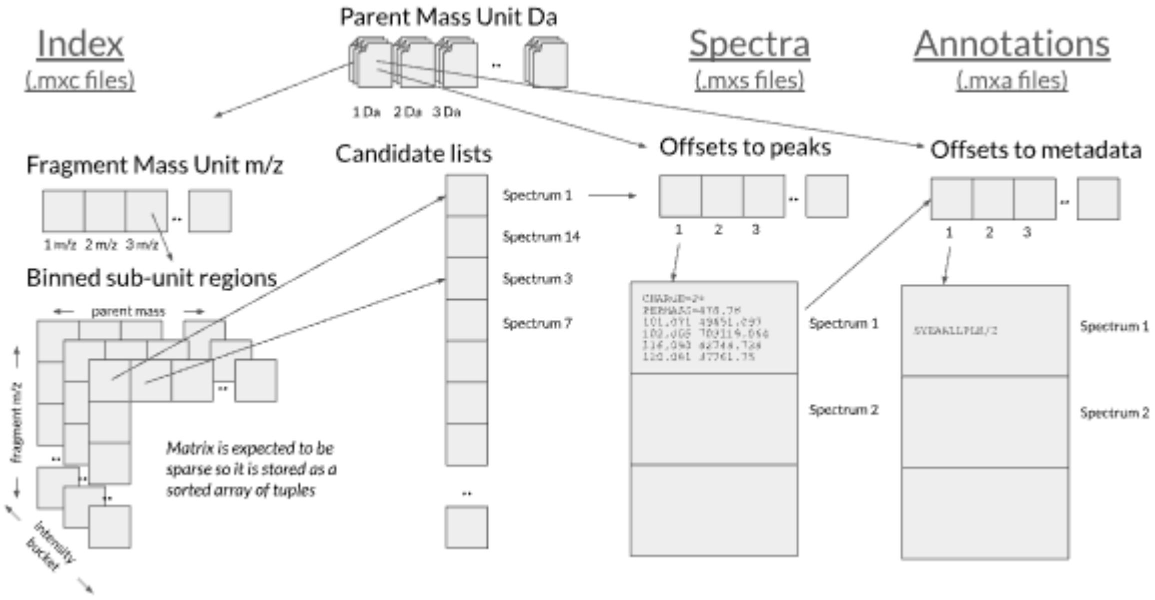


Figure 4.4. The overview of the peak indexing data structures. For each parent mass unit Da, there are three files, an index file (.mxc), a spectrum file (.mxs), and an annotation file (.mxa). The process for moving from a single peak, at some parent mass, fragment mass, and intensity to all the spectra containing that peak, and subsequently all annotations for these spectra are considered below. To index the peak, first the unit fragment mass points to a sparse 3-dimensional matrix with the binned sub-unit (fractional part, between the current mass m, inclusive, and the next mass $m+1$, exclusive) fragment mass, parent mass, and intensity. That then points to an index offset in the candidates where all the candidates containing that peak are listed. From there, the spectra and annotations can be directly indexed by the previously found spectrum peak id.

Table 4.1. Library exact searches help to determine the identity of of the input MS/MS spectrum Library analog searches help to determine the what identifications are similar to the input spectrum, and at what delta mass they are found. Repository exact searches: find all matching MS/MS spectra and the samples/tissues/datasets where they occur and confirm consistency of identifications across all datasets. Repository analog searches: find all variants of a peptide/molecule and all datasets where they occur.

		Median (s)	***St. Dev. (s)***
MassIVE-KB (Library)	Exact	0.32s	0.01s
	Analog	0.78s	0.33s
GNPS (Library)	Exact	0.30s	0.02s
	Analog	0.38s	0.05s
MassIVE (Repository)	Exact	0.48s	0.13s
	Analog	56s	26s
GNPS (Repository)	Exact	0.69s	0.67s
	Analog	23s	19s

Bibliography

[1] Yaoyang Zhang, Bryan R Fonslow, Bing Shan, Moon-Chang Baek, and John R Yates. "Protein analysis by shotgun/bottom-up proteomics." In: *Chemical reviews* 113.4 (Apr. 2013), pp. 2343–94. DOI: 10.1021/cr3003533. URL: http://www.ncbi.nlm.nih.gov/pubmed/23438204.

[2] Ariel Bensimon, Albert J R Heck, and Ruedi Aebersold. "Mass spectrometry-based proteomics and network biology." In: *Annual review of biochemistry* 81 (2012), pp. 379–405. DOI: 10.1146/annurev-biochem-072909-100424. URL: http://www.ncbi.nlm.nih.gov/pubmed/22439968.

[3] Gilbert S Omenn, Lydie Lane, Emma K Lundberg, Christopher M Overall, and Eric W Deutsch. "Progress on the HUPO Draft Human Proteome: 2017 Metrics of the Human Proteome Project." In: *Journal of proteome research* 16.12 (Dec. 2017), pp. 4281–4287. DOI: 10.1021/acs.jproteome.7b00375. URL: http://www.ncbi.nlm.nih.gov/pubmed/28853897.

[4] Mingxun Wang, Jian Wang, Jeremy Carver, Benjamin S. Pullman, Seong Won Cha, and Nuno Bandeira. "Assembling the Community-Scale Discoverable Human Proteome." In: *Cell systems* 7.4 (2018), 412–421.e5. DOI: 10.1016/j.cels.2018.08.004. URL: http://www.ncbi.nlm.nih.gov/pubmed/30172843.

[5] Daniel P Zolg, Mathias Wilhelm, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Bernard Delanghe, Derek J Bailey, Siegfried Gessulat, Hans-Christian Ehrlich, Maximilian Weininger, Peng Yu, Judith Schlegl, Karl Kramer, Tobias Schmidt, Ulrike Kusebauch, Eric W Deutsch, Ruedi Aebersold, Robert L Moritz, Holger Wenschuh, Thomas Moehring, Stephan Aiche, Andreas Huhmer, Ulf Reimer, and Bernhard Kuster. "Building ProteomeTools based on a complete synthetic human proteome." In: *Nature methods* 14.3 (2017), pp. 259–262. DOI: 10.1038/nmeth.4153. URL: http://www.ncbi.nlm.nih.gov/pubmed/28135259.

[6] Edward L Huttlin, Raphael J Bruckner, Joao A Paulo, Joe R Cannon, Lily Ting, Kurt Baltier, Greg Colby, Fana Gebreab, Melanie P Gygi, Hannah Parzen, John Szpyt, Stanley Tam, Gabriela Zarraga, Laura Pontano-Vaites, Sharan Swarup, Anne E White, Devin K Schweppe, Ramin Rad, Brian K Erickson, Robert A Obar, K G Guruharsha, Kejie Li, Spyros Artavanis-Tsakonas, Steven P Gygi, and J Wade Harper. "Architecture of the human interactome defines protein communities and disease networks." In: *Nature* 545.7655 (2017), pp. 505–509. DOI: 10.1038/nature22366. URL: http://www.ncbi.nlm.nih.gov/pubmed/28514442.

[7] Yang Liu, Wantao Ying, Zhe Ren, Wei Gu, Yang Zhang, Guoquan Yan, Pengyuan Yang, Yinkun Liu, Xuefei Yin, Cheng Chang, Jing Jiang, Fengxu Fan, Chengpu Zhang, Ping Xu, Quanhui Wang, Bo Wen, Liang Lin, Tingyou Wang, Chaoqin Du, Jiayong Zhong, Tong Wang, Qing-Yu He, Xiaohong Qian, Xiaomin Lou, Gong Zhang, and Fan Zhong. "Chromosome-8-coded proteome of Chinese Chromosome Proteome Data set (CCPD) 2.0 with partial immunohistochemical verifications." In: *Journal of proteome research* 13.1 (Jan. 2014), pp. 126–36. DOI: 10.1021/pr400902u. URL: http://www.ncbi.nlm.nih.gov/pubmed/24328083.

[8] Min-Sik Kim, Sneha M Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S Manda, Raghothama Chaerkady, Anil K Madugundu, Dhanashree S Kelkar, Ruth Isserlin, Shobhit Jain, Joji K Thomas, Babylakshmi Muthusamy, Pamela Leal-Rojas, Praveen Kumar, Nandini A Sahasrabuddhe, Lavanya Balakrishnan, Jayshree Advani, Bijesh George, Santosh Renuse, Lakshmi Dhevi N Selvan, Arun H Patil, Vishalakshi Nanjappa, Aneesha Radhakrishnan, Samarjeet Prasad, Tejaswini Subbannayya, Rajesh Raju, Manish Kumar, Sreelakshmi K Sreenivasamurthy, Arivusudar Marimuthu, Gajanan J Sathe, Sandip Chavan, Keshava K Datta, Yashwanth Subbannayya, Apeksha Sahu, Soujanya D Yelamanchi, Savita Jayaram, Pavithra Rajagopalan, Jyoti Sharma, Krishna R Murthy, Nazia Syed, Renu Goel, Aafaque A Khan, Sartaj Ahmad, Gourav Dey, Keshav Mudgal, Aditi Chatterjee, Tai-Chung Huang, Jun Zhong, Xinyan Wu, Patrick G Shaw, Donald Freed, Muhammad S Zahari, Kanchan K Mukherjee, Susarla Krishna Subramanian Shankar, Anita Mahadevan, Henry Lam, Christopher J Mitchell, Susarla Krishna Subramanian Shankar, Parthasarathy Satishchandra, John T Schroeder, Ravi Sirdeshmukh, Anirban Maitra, Steven D Leach, Charles G Drake, Marc K Halushka, T S Keshava Prasad, Ralph H Hruban, Candace L Kerr, Gary D Bader, Christine A Iacobuzio-Donahue, Harsha Gowda, and Akhilesh Pandey. "A draft map of the human proteome." In: *Nature* 509.7502 (May 2014), pp. 575–81. DOI: 10.1038/nature13302. URL: http://www.ncbi.nlm.nih.gov/pubmed/24870542.

[9] Kirti Sharma, Rochelle C J D'Souza, Stefka Tyanova, Christoph Schaab, Jacek R Wiśniewski, Jürgen Cox, and Matthias Mann. "Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling." In: *Cell reports* 8.5 (Sept. 2014), pp. 1583–94. DOI: 10.1016/j.celrep.2014.07.036. URL: http://www.ncbi.nlm.nih.gov/pubmed/25159151.

[10] Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, Toby Mathieson, Simone Lemeer, Karsten Schnatbaum, Ulf Reimer, Holger Wenschuh, Martin Mollenhauer, Julia Slotta-Huspenina, Joos-Hendrik Boese, Marcus Bantscheff, Anja Gerstmair, Franz Faerber, and Bernhard Kuster. "Mass-spectrometry-based draft of the human proteome." In: *Nature* 509.7502 (May 2014), pp. 582–7. DOI: 10.1038/nature13319. URL: http://www.ncbi.nlm.nih.gov/pubmed/24870543.

[11] Yves Vandenbrouck, Lydie Lane, Christine Carapito, Paula Duek, Karine Rondel, Christophe Bruley, Charlotte Macron, Anne Gonzalez de Peredo, Yohann Couté, Karima Chaoui, Emmanuelle Com, Alain Gateau, Anne-Marie Hesse, Marlene Marcellin, Loren Méar, Emmanuelle Mouton-Barbosa, Thibault Robin, Odile Burlet-Schiltz, Sarah Cianferani, Myr-

iam Ferro, Thomas Fréour, Cecilia Lindskog, Jérôme Garin, and Charles Pineau. “Looking for Missing Proteins in the Proteome of Human Spermatozoa: An Update.” In: *Journal of proteome research* 15.11 (2016), pp. 3998–4019. DOI: 10.1021/acs.jproteome.6b00400. URL: http://www.ncbi.nlm.nih.gov/pubmed/27444420.

[12] Sangtae Kim and Pavel A Pevzner. “MS-GF+ makes progress towards a universal database search tool for proteomics.” In: *Nature communications* 5 (Oct. 2014), pp. 5277–5277. DOI: 10.1038/ncomms6277. URL: http://www.ncbi.nlm.nih.gov/pubmed/25358478.

[13] Joshua E Elias and Steven P Gygi. “Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry”. In: *Nature methods* 4.3 (Mar. 2007), pp. 207–214. DOI: 10.1038/nmeth1019. URL: http://www.ncbi.nlm.nih.gov/pubmed/17327847.

[14] Nitin Gupta and Pavel A Pevzner. “False discovery rates of protein identifications: a strike against the two-peptide rule”. In: *J. Proteome Res.* 8.9 (Sept. 2009), pp. 4173–4181. DOI: 10.1021/pr9004794. URL: http://www.ncbi.nlm.nih.gov/pubmed/19627159.

[15] Mikhail M Savitski, Mathias Wilhelm, Hannes Hahne, Bernhard Kuster, and Marcus Bantscheff. “A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets.” In: *Molecular & cellular proteomics* 14.9 (Sept. 2015), pp. 2394–404. DOI: 10.1074/mcp.M114.046995. URL: http://www.ncbi.nlm.nih.gov/pubmed/25987413.

[16] Peter V Hornbeck, Bin Zhang, Beth Murray, Jon M Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. “PhosphoSitePlus, 2014: mutations, PTMs and recalibrations.” In: *Nucleic acids research* 43 (Database issue Jan. 2015), pp. D512–20. DOI: 10.1093/nar/gku1267.

[17] The UniProt Consortium. “UniProt: the universal protein knowledgebase.” In: *Nucleic acids research* 45 (D1 Jan. 2017), pp. D158–D169. DOI: 10.1093/nar/gkw1099.

[18] Pascale Gaudet, Pierre-André Michel, Monique Zahn-Zabal, Aurore Britan, Isabelle Cusin, Marcin Domagalski, Paula D Duek, Alain Gateau, Anne Gleizes, Valérie Hinard, Valentine Rech de Laval, JinJin Lin, Frederic Nikitin, Mathieu Schaeffer, Daniel Teixeira, Lydie Lane, and Amos Bairoch. “The neXtProt knowledgebase on human proteins: 2017 update.” In: *Nucleic acids research* 45 (D1 Jan. 2017), pp. D177–D182. DOI: 10.1093/nar/gkw1062.

[19] Bronwen L Aken, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Friederike Bernsdorff, Jyothish Bhai, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E Hunt, Sophie H Janacek, Thomas Juettemann, Stephen Keenan, Matthew R Laird, Ilias Lavidas, Thomas Maurel, William McLaren, Benjamin Moore, Daniel N Murphy, Rishi Nag, Victoria Newman, Michael Nuhn, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Daniel Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Steven P Wilder, Amonida Zadissa, Myrto Kostadima, Fergal J Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Daniel M Staines, Stephen J Trevanion, Fiona Cunningham, Andrew Yates, Daniel R Zerbino, and

Paul Flicek. “Ensembl 2017.” In: *Nucleic acids research* 45 (D1 Jan. 2017), pp. D635–D642. DOI: 10.1093/nar/gkw1104. URL: http://www.ncbi.nlm.nih.gov/pubmed/27899575.

[20] Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G Izuogu, Sophie H Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R Laird, Ilias Lavidas, Zhicheng Liu, Jane E Loveland, Thomas Maurel, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M Staines, Stephen J Trevanion, Bronwen L Aken, Fiona Cunningham, Andrew Yates, and Paul Flicek. “Ensembl 2018.” In: *Nucleic acids research* 46 (D1 Jan. 2018), pp. D754–D761. DOI: 10.1093/nar/gkx1098. URL: http://www.ncbi.nlm.nih.gov/pubmed/29155950.

[21] Jan Odvárko. *jscolor*. E-mail. 2018. URL: http://jscolor.com.

[22] Eric W Deutsch, Christopher M Overall, Jennifer E Van Eyk, Mark S Baker, Young-Ki Paik, Susan T Weintraub, Lydie Lane, Lennart Martens, Yves Vandenbrouck, Ulrike Kusebauch, William S Hancock, Henning Hermjakob, Ruedi Aebersold, Robert L Moritz, and Gilbert S Omenn. “Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1.” In: *Journal of proteome research* 15.11 (Nov. 2016), pp. 3961–3970. DOI: 10.1021/acs.jproteome.6b00392. URL: http://www.ncbi.nlm.nih.gov/pubmed/27490519.

[23] Sunghee Woo, Seong Won Cha, Gennifer Merrihew, Yupeng He, Natalie Castellana, Clark Guest, Michael MacCoss, and Vineet Bafna. “Proteogenomic database construction driven from large scale RNA-seq data.” In: *Journal of proteome research* 13.1 (Jan. 2014), pp. 21–8. DOI: 10.1021/pr400294c. URL: http://www.ncbi.nlm.nih.gov/pubmed/23802565.

[24] Sunghee Woo, Seong Won Cha, Stefano Bonissone, Seungjin Na, David L Tabb, Pavel A Pevzner, and Vineet Bafna. “Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer.” In: *Journal of proteome research* 14.9 (Sept. 2015), pp. 3555–67. DOI: 10.1021/acs.jproteome.5b00264. URL: http://www.ncbi.nlm.nih.gov/pubmed/26139413.

[25] Paul A Rudnick, Sanford P Markey, Jeri Roth, Yuri Mirokhin, Xinjian Yan, Dmitrii V Tchekhovskoi, Nathan J Edwards, Ratna R Thangudu, Karen A Ketchum, Christopher R Kinsinger, Mehdi Mesri, Henry Rodriguez, and Stephen E Stein. “A Description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) Common Data Analysis Pipeline.” In: *Journal of proteome research* 15.3 (Mar. 2016), pp. 1023–32. DOI: 10.1021/acs.jproteome.5b01091. URL: http://www.ncbi.nlm.nih.gov/pubmed/26860878.

[26] Dongxue Wang, Basak Eraslan, Thomas Wieland, Björn Hallström, Thomas Hopf, Daniel Paul Zolg, Jana Zecha, Anna Asplund, Li-Hua Li, Chen Meng, Martin Frejno, Tobias Schmidt, Karsten Schnatbaum, Mathias Wilhelm, Frederik Ponten, Mathias Uhlen, Julien Gagneur, Hannes Hahne, and Bernhard Kuster. “A deep proteome and transcriptome

abundance atlas of 29 healthy human tissues". In: *Molecular Systems Biology* 15.2 (Feb. 18, 2019), e8503. ISSN: 1744-4292. DOI: 10.15252/msb.20188503.

[27] Lihua Jiang, Meng Wang, Shin Lin, Ruiqi Jian, Xiao Li, Joanne Chan, Guanlan Dong, Huaying Fang, Aaron E. Robinson, GTEx Consortium, and Michael P. Snyder. "A Quantitative Proteome Map of the Human Body". In: *Cell* 183.1 (Oct. 1, 2020), 269–283.e19. ISSN: 1097-4172. DOI: 10.1016/j.cell.2020.08.036.

[28] Andrew Keller, Alexey I. Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search". In: *Analytical Chemistry* 74.20 (Oct. 15, 2002), pp. 5383–5392. ISSN: 0003-2700. DOI: 10.1021/ac025747h.

[29] Lukas Käll, Jesse D. Canterbury, Jason Weston, William Stafford Noble, and Michael J. MacCoss. "Semi-supervised learning for peptide identification from shotgun proteomics datasets". In: *Nature Methods* 4.11 (Nov. 2007), pp. 923–925. ISSN: 1548-7091. DOI: 10.1038/nmeth1113.

[30] Mathias Walzer, Lucia Espona Pernas, Sara Nasso, Wout Bittremieux, Sven Nahnsen, Pieter Kelchtermans, Peter Pichler, Henk W P van den Toorn, An Staes, Jonathan Vandenbussche, Michael Mazanek, Thomas Taus, Richard A Scheltema, Christian D Kelstrup, Laurent Gatto, Bas van Breukelen, Stephan Aiche, Dirk Valkenborg, Kris Laukens, Kathryn S Lilley, Jesper V Olsen, Albert J R Heck, Karl Mechtler, Ruedi Aebersold, Kris Gevaert, Juan Antonio Vizcaíno, Henning Hermjakob, Oliver Kohlbacher, and Lennart Martens. "qcML: an exchange format for quality control metrics from mass spectrometry experiments." In: *Molecular & cellular proteomics* 13.8 (Aug. 2014), pp. 1905–13. DOI: 10.1074/mcp.M113.035907. URL: http://www.ncbi.nlm.nih.gov/pubmed/24760958.

[31] Mathieu Schaeffer, Alain Gateau, Daniel Teixeira, Pierre-André Michel, Monique Zahn-Zabal, and Lydie Lane. "The neXtProt peptide uniqueness checker: a tool for the proteomics community". In: *Bioinformatics (Oxford, England)* 33.21 (Nov. 1, 2017), pp. 3471–3472. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btx318.

[32] Benjamin S Pullman, Julie Wertz, Jeremy Carver, and Nuno Bandeira. "ProteinExplorer: A Repository-Scale Resource for Exploration of Protein Detection in Public Mass Spectrometry Data Sets." In: *Journal of proteome research* 17.12 (2018), pp. 4227–4234. DOI: 10.1021/acs.jproteome.8b00496. URL: http://www.ncbi.nlm.nih.gov/pubmed/30985146.

[33] Johannes Griss, Andrew R Jones, Timo Sachsenberg, Mathias Walzer, Laurent Gatto, Jürgen Hartler, Gerhard G Thallinger, Reza M Salek, Christoph Steinbeck, Nadin Neuhauser, Jürgen Cox, Steffen Neumann, Jun Fan, Florian Reisinger, Qing-Wei Xu, Noemi Del Toro, Yasset Pérez-Riverol, Fawaz Ghali, Nuno Bandeira, Ioannis Xenarios, Oliver Kohlbacher, Juan Antonio Vizcaíno, and Henning Hermjakob. "The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience." In: *Molecular & cellular proteomics* 13.10 (Oct. 2014), pp. 2765–75. DOI: 10.1074/mcp.O113.036681. URL: http://www.ncbi.nlm.nih.gov/pubmed/24980485.

[34] Andrew R Jones, Martin Eisenacher, Gerhard Mayer, Oliver Kohlbacher, Jennifer Siepen, Simon J Hubbard, Julian N Selley, Brian C Searle, James Shofstahl, Sean L Seymour, Randall Julian, Pierre-Alain Binz, Eric W Deutsch, Henning Hermjakob, Florian Reisinger, Johannes Griss, Juan Antonio Vizcaíno, Matthew Chambers, Angel Pizarro, and David Creasy. "The mzIdentML data standard for mass spectrometry-based proteomics results." In: *Molecular & cellular proteomics* 11.7 (July 2012), pp. M111.014381–M111.014381. DOI: 10.1074/mcp.M111.014381. URL: http://www.ncbi.nlm.nih.gov/pubmed/22375074.

[35] UniProt Consortium. "UniProt: the universal protein knowledgebase in 2021". In: *Nucleic Acids Research* 49 (D1 Jan. 8, 2021), pp. D480–D489. ISSN: 1362-4962. DOI: 10.1093/nar/gkaa1100.

[36] Eric W Deutsch, Lydie Lane, Christopher M Overall, Nuno Bandeira, Mark S Baker, Charles Pineau, Robert L Moritz, Fernando Corrales, Sandra Orchard, Jennifer E Van Eyk, Young-Ki Paik, Susan T Weintraub, Yves Vandenbrouck, and Gilbert S Omenn. "Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0." In: *Journal of proteome research* 18.12 (2019), pp. 4108–4116. DOI: 10.1021/acs.jproteome.9b00542. URL: http://www.ncbi.nlm.nih.gov/pubmed/31599596.

[37] Andrew Thompson, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Schmidt, Thomas Neumann, R. Johnstone, A. Karim A. Mohammed, and Christian Hamon. "Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS". In: *Analytical Chemistry* 75.8 (Apr. 15, 2003), pp. 1895–1904. ISSN: 0003-2700. DOI: 10.1021/ac0262560.

[38] Sebastian Wiese, Kai A. Reidegeld, Helmut E. Meyer, and Bettina Warscheid. "Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research". In: *Proteomics* 7.3 (Feb. 2007), pp. 340–350. ISSN: 1615-9853. DOI: 10.1002/pmic.200600422.

[39] Adrian Guthals, Karl R Clauser, Ari M Frank, and Nuno Bandeira. "Sequencing-grade de novo analysis of MS/MS triplets (CID/HCD/ETD) from overlapping peptides." In: *Journal of proteome research* 12.6 (June 2013), pp. 2846–57. DOI: 10.1021/pr400173d. URL: http://www.ncbi.nlm.nih.gov/pubmed/23679345.

[40] David L Tabb, Michael J MacCoss, Christine C Wu, Scott D Anderson, and John R Yates. "Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility." In: *Analytical chemistry* 75.10 (May 2003), pp. 2470–7. URL: http://www.ncbi.nlm.nih.gov/pubmed/12918992.

[41] Eric W Deutsch, Nuno Bandeira, Vagisha Sharma, Yasset Perez-Riverol, Jeremy J Carver, Deepti J Kundu, David García-Seisdedos, Andrew F Jarnuczak, Suresh Hewapathirana, Benjamin S Pullman, Julie Wertz, Zhi Sun, Shin Kawano, Shujiro Okuda, Yu Watanabe, Henning Hermjakob, Brendan MacLean, Michael J MacCoss, Yunping Zhu, Yasushi Ishihama, and Juan A Vizcaíno. "The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics." In: *Nucleic acids research* 48 (D1 2020), pp. D1145–D1152. DOI: 10.1093/nar/gkz984. URL: http://www.ncbi.nlm.nih.gov/pubmed/31686107.

[42] Dorte B Bekker-Jensen, Christian D Kelstrup, Tanveer S Batth, Sara C Larsen, Christa Haldrup, Jesper B Bramsen, Karina D Sørensen, Søren Høyer, Torben F Ørntoft, Claus L Andersen, Michael L Nielsen, and Jesper V Olsen. “An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes.” In: *Cell systems* 4.6 (2017), 587–599.e4. DOI: 10.1016/j.cels.2017.05.009. URL: http://www.ncbi.nlm.nih.gov/pubmed/28601559.

[43] Heeyoun Hwang, Ji Eun Jeong, Hyun Kyoung Lee, Ki Na Yun, Hyun Joo An, Bonghee Lee, Young-Ki Paik, Tae Seok Jeong, Gi Taek Yee, Jin Young Kim, and Jong Shin Yoo. “Identification of Missing Proteins in Human Olfactory Epithelial Tissue by Liquid Chromatography-Tandem Mass Spectrometry”. In: *Journal of Proteome Research* 17.12 (Dec. 7, 2018), pp. 4320–4324. ISSN: 1535-3907. DOI: 10.1021/acs.jproteome.8b00408.

[44] Mark S. Baker, Seong Beom Ahn, Abidali Mohamedali, Mohammad T. Islam, David Cantor, Peter D. Verhaert, Susan Fanayan, Samridhi Sharma, Edouard C. Nice, Mark Connor, and Shoba Ranganathan. “Accelerating the search for the missing proteins in the human proteome”. In: *Nature Communications* 8 (Jan. 24, 2017), p. 14271. ISSN: 2041-1723. DOI: 10.1038/ncomms14271.

[45] Robert J Chalkley and Karl R Clauser. “Modification site localization scoring: strategies and performance.” In: *Molecular & cellular proteomics* 11.5 (May 2012), pp. 3–14. DOI: 10.1074/mcp.R111.015305. URL: http://www.ncbi.nlm.nih.gov/pubmed/22328712.

[46] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. “UpSet: Visualization of Intersecting Sets”. In: *IEEE transactions on visualization and computer graphics* 20.12 (Dec. 2014), pp. 1983–1992. ISSN: 1941-0506. DOI: 10.1109/TVCG.2014.2346248.

[47] Eric W. Deutsch, Yasset Perez-Riverol, Jeremy Carver, Shin Kawano, Luis Mendoza, Tim Van Den Bossche, Ralf Gabriels, Pierre-Alain Binz, Benjamin Pullman, Zhi Sun, Jim Shofstahl, Wout Bittremieux, Tytus D. Mak, Joshua Klein, Yunping Zhu, Henry Lam, Juan Antonio Vizcaíno, and Nuno Bandeira. “Universal Spectrum Identifier for mass spectra”. In: *Nature Methods* 18.7 (July 2021), pp. 768–770. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01184-6.

[48] Ruedi Aebersold and Matthias Mann. “Mass spectrometry-based proteomics”. In: *Nature* 422.6928 (Mar. 13, 2003), pp. 198–207. ISSN: 0028-0836. DOI: 10.1038/nature01511.

[49] Nathan J Edwards, Mauricio Oberti, Ratna R Thangudu, Shuang Cai, Peter B McGarvey, Shine Jacob, Subha Madhavan, and Karen A Ketchum. “The CPTAC Data Portal: A Resource for Cancer Proteomics Research.” In: *Journal of proteome research* 14.6 (June 2015), pp. 2707–13. DOI: 10.1021/pr501254j. URL: http://www.ncbi.nlm.nih.gov/pubmed/25873244.

[50] Bo Shen, Xiao Yi, Yaoting Sun, Xiaojie Bi, Juping Du, Chao Zhang, Sheng Quan, Fangfei Zhang, Rui Sun, Liujia Qian, Weigang Ge, Wei Liu, Shuang Liang, Hao Chen, Ying Zhang, Jun Li, Jiaqin Xu, Zebao He, Baofu Chen, Jing Wang, Haixi Yan, Yufen Zheng, Donglian Wang, Jiansheng Zhu, Ziqing Kong, Zhouyang Kang, Xiao Liang, Xuan Ding, Guan Ruan, Nan Xiang, Xue Cai, Huanhuan Gao, Lu Li, Sainan Li, Qi Xiao, Tian Lu,

Yi Zhu, Huafen Liu, Haixiao Chen, and Tiannan Guo. “Proteomic and Metabolomic Characterization of COVID-19 Patient Sera”. In: *Cell* 182.1 (July 9, 2020), 59–72.e15. ISSN: 1097-4172. DOI: 10.1016/j.cell.2020.05.032.

[51] Susan Klaeger, Stephanie Heinzlmeir, Mathias Wilhelm, Harald Polzer, Binje Vick, Paul-Albert Koenig, Maria Reinecke, Benjamin Ruprecht, Svenja Petzoldt, Chen Meng, Jana Zecha, Katrin Reiter, Huichao Qiao, Dominic Helm, Heiner Koch, Melanie Schoof, Giulia Canevari, Elena Casale, Stefania Re Depaolini, Annette Feuchtinger, Zhixiang Wu, Tobias Schmidt, Lars Rueckert, Wilhelm Becker, Jan Huenges, Anne-Kathrin Garz, Bjoern-Oliver Gohlke, Daniel Paul Zolg, Gian Kayser, Tonu Vooder, Robert Preissner, Hannes Hahne, Neeme Tõnisson, Karl Kramer, Katharina Götze, Florian Bassermann, Judith Schlegl, Hans-Christian Ehrlich, Stephan Aiche, Axel Walch, Philipp A. Greif, Sabine Schneider, Eduard Rudolf Felder, Juergen Ruland, Guillaume Médard, Irmela Jeremias, Karsten Spiekermann, and Bernhard Kuster. “The target landscape of clinical kinase drugs”. In: *Science (New York, N.Y.)* 358.6367 (Dec. 1, 2017), eaan4368. ISSN: 1095-9203. DOI: 10.1126/science.aan4368.

[52] Ari M. Frank, Nuno Bandeira, Zhouxin Shen, Stephen Tanner, Steven P. Briggs, Richard D. Smith, and Pavel A. Pevzner. “Clustering millions of tandem mass spectra”. In: *Journal of Proteome Research* 7.1 (Jan. 2008), pp. 113–122. ISSN: 1535-3893 (Print)\r1535-3893. DOI: 10.1021/pr070361e. URL: http://www.ncbi.nlm.nih.gov/pubmed/18067247.

[53] Ari M Frank, Matthew E Monroe, Anuj R Shah, Jeremy J Carver, Nuno Bandeira, Ronald J Moore, Gordon A Anderson, Richard D Smith, and Pavel A Pevzner. “Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra”. In: *Nature Methods* 8.7 (May 2011), pp. 587–591. ISSN: 1548-7105 (Electronic)\r1548-7091 (Linking). DOI: 10.1038/nmeth.1609. URL: http://www.nature.com/doifinder/10.1038/nmeth.1609.

[54] Johannes Griss, Joseph M Foster, Henning Hermjakob, and Juan Antonio Vizcaíno. “PRIDE Cluster: building a consensus of proteomics data.” In: *Nature methods* 10.2 (Feb. 2013), pp. 95–6. DOI: 10.1038/nmeth.2343. URL: http://www.ncbi.nlm.nih.gov/pubmed/23361086.

[55] Johannes Griss, Yasset Perez-Riverol, Steve Lewis, David L Tabb, José A Dianes, Noemi Del-Toro, Marc Rurik, Mathias W Walzer, Oliver Kohlbacher, Henning Hermjakob, Rui Wang, and Juan Antonio Vizcaíno. “Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets.” In: *Nature methods* 13.8 (Aug. 2016), pp. 651–656. DOI: 10.1038/nmeth.3902. URL: http://www.ncbi.nlm.nih.gov/pubmed/27493588.

[56] Matthew The and Lukas Käll. “MaRaCluster: A Fragment Rarity Metric for Clustering Fragment Spectra in Shotgun Proteomics.” In: *Journal of proteome research* 15.3 (Mar. 2016), pp. 713–20. DOI: 10.1021/acs.jproteome.5b00749. URL: http://www.ncbi.nlm.nih.gov/pubmed/26653874.

[57] Wout Bittremieux, Damon H. May, Jeffrey Bilmes, and William Stafford Noble. "A learned embedding for efficient joint analysis of millions of mass spectra". In: *Nature Methods* 19.6 (June 2022), pp. 675–678. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01496-1.

[58] Jian Wang, Josué Pérez-Santiago, Jonathan E Katz, Parag Mallick, and Nuno Bandeira. "Peptide identification from mixture tandem mass spectra." In: *Molecular & cellular proteomics* 9.7 (July 2010), pp. 1476–85. DOI: 10.1074/mcp.M000136-MCP201. URL: http://www.ncbi.nlm.nih.gov/pubmed/20348588.

[59] Nuno Bandeira, Dekel Tsur, Ari Frank, and Pavel A Pevzner. "Protein identification by spectral networks analysis." In: *Proceedings of the National Academy of Sciences of the United States of America* 104.15 (Apr. 2007), pp. 6140–5. DOI: 10.1073/pnas.0701130104. URL: http://www.ncbi.nlm.nih.gov/pubmed/17404225.

[60] Jeramie Watrous, Patrick Roach, Theodore Alexandrov, Brandi S Heath, Jane Y Yang, Roland D Kersten, Menno van der Voort, Kit Pogliano, Harald Gross, Jos M Raaijmakers, Bradley S Moore, Julia Laskin, Nuno Bandeira, and Pieter C Dorrestein. "Mass spectral molecular networking of living microbial colonies." In: *Proceedings of the National Academy of Sciences of the United States of America* 109.26 (June 2012), E1743–52. DOI: 10.1073/pnas.1203689109. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1203689109.

[61] T. F. Smith and M. S. Waterman. "Identification of common molecular subsequences". In: *Journal of Molecular Biology* 147.1 (Mar. 25, 1981), pp. 195–197. ISSN: 0022-2836. DOI: 10.1016/0022-2836(81)90087-5.

[62] Seung-Ah Lee, Kryscilla Jian Zhang Yang, Pierre-Jacques Brun, Josie A. Silvaroli, Jason J. Yuen, Igor Shmarakov, Hongfeng Jiang, Jun B. Feranil, Xueting Li, Atreju I. Lackey, Wojciech Krężel, Rudolph L. Leibel, Jenny Libien, Judith Storch, Marcin Golczak, and William S. Blaner. "Retinol-binding protein 2 (RBP2) binds monoacylglycerols and modulates gut endocrine signaling and body weight". In: *Science Advances* 6.11 (Mar. 2020), eaay8937. ISSN: 2375-2548. DOI: 10.1126/sciadv.aay8937. URL: https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.aay8937 (visited on 05/14/2021).

[63] Mathias Uhlén, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, IngMarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-Khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-Henrik Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M. Schwenk, Marica Hamsten, Kalle von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar von Heijne, Jens Nielsen, and Fredrik Pontén. "Proteomics. Tissue-based map of the human proteome". In: *Science (New York, N.Y.)* 347.6220 (Jan. 23, 2015), p. 1260419. ISSN: 1095-9203. DOI: 10.1126/science.1260419.

[64] Xiangrong Li and Su Wang. "Binding of glutathione and melatonin to human serum albumin: a comparative study". In: *Colloids and Surfaces. B, Biointerfaces* 125 (Jan. 1, 2015), pp. 96–103. ISSN: 1873-4367. DOI: 10.1016/j.colsurfb.2014.11.023.

[65] Maria Monica Castellanos and Coray M. Colina. “Molecular dynamics simulations of human serum albumin and role of disulfide bonds”. In: *The Journal of Physical Chemistry. B* 117.40 (Oct. 10, 2013), pp. 11895–11905. ISSN: 1520-5207. DOI: 10.1021/jp402994r.

[66] James A. Yergey. “A general approach to calculating isotopic distributions for mass spectrometry”. In: *Journal of mass spectrometry: JMS* 55.8 (Aug. 2020), e4498. ISSN: 1096-9888. DOI: 10.1002/jms.4498.

[67] Stephane Houel, Robert Abernathy, Kutralanathan Renganathan, Karen Meyer-Arendt, Natalie G. Ahn, and William M. Old. “Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies”. In: *Journal of Proteome Research* 9.8 (Aug. 6, 2010), pp. 4152–4160. ISSN: 1535-3907. DOI: 10.1021/pr1003856.

[68] Mingxun Wang, Alan K Jarmusch, Fernando Vargas, Alexander A Aksenov, Julia M Gauglitz, Kelly Weldon, Daniel Petras, Ricardo da Silva, Robert Quinn, Alexey V Melnik, Justin J J van der Hooft, Andrés Mauricio Caraballo-Rodríguez, Louis Felix Nothias, Christine M Aceves, Morgan Panitchpakdi, Elizabeth Brown, Francesca Di Ottavio, Nicole Sikora, Emmanuel O Elijah, Lara Labarta-Bajo, Emily C Gentry, Shabnam Shalapour, Kathleen E Kyle, Sara P Puckett, Jeramie D Watrous, Carolina S Carpenter, Amina Bouslimani, Madeleine Ernst, Austin D Swafford, Elina I Zúñiga, Marcy J Balunas, Jonathan L Klassen, Rohit Loomba, Rob Knight, Nuno Bandeira, and Pieter C Dorrestein. “Mass spectrometry searches using MASST.” In: *Nature biotechnology* 38.1 (2020), pp. 23–26. DOI: 10.1038/s41587-019-0375-9. URL: http://www.ncbi.nlm.nih.gov/pubmed/31894142.

[69] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, Carla Porto, Amina Bouslimani, Alexey V Melnik, Michael J Meehan, Wei-Ting Liu, Max Crüsemann, Paul D Boudreau, Eduardo Esquenazi, Mario Sandoval-Calderón, Roland D Kersten, Laura A Pace, Robert A Quinn, Katherine R Duncan, Cheng-Chih Hsu, Dimitrios J Floros, Ronnie G Gavilan, Karin Kleigrewe, Trent Northen, Rachel J Dutton, Delphine Parrot, Erin E Carlson, Bertrand Aigle, Charlotte F Michelsen, Lars Jelsbak, Christian Sohlenkamp, Pavel Pevzner, Anna Edlund, Jeffrey McLean, Jörn Piel, Brian T Murphy, Lena Gerwick, Chih-Chuang Liaw, Yu-Liang Yang, Hans-Ulrich Humpf, Maria Maansson, Robert A Keyzers, Amy C Sims, Andrew R Johnson, Ashley M Sidebottom, Brian E Sedio, Andreas Klitgaard, Charles B Larson, Cristopher A Boya P, Daniel Torres-Mendoza, David J Gonzalez, Denise B Silva, Lucas M Marques, Daniel P Demarque, Egle Pociute, Ellis C O’Neill, Enora Briand, Eric J N Helfrich, Eve A Granatosky, Evgenia Glukhov, Florian Ryffel, Hailey Houson, Hosein Mohimani, Jenan J Kharbush, Yi Zeng, Julia A Vorholt, Kenji L Kurita, Pep Charusanti, Kerry L McPhail, Kristian Fog Nielsen, Lisa Vuong, Maryam Elfeki, Matthew F Traxler, Niclas Engene, Nobuhiro Koyama, Oliver B Vining, Ralph Baric, Ricardo R Silva, Samantha J Mascuch, Sophie Tomasi, Stefan Jenkins, Venkat Macherla, Thomas Hoffman, Vinayak Agarwal, Philip G Williams, Jingqui Dai, Ram Neupane, Joshua Gurr, Andrés M C Rodríguez, Anne Lamsa, Chen Zhang, Kathleen Dorrestein, Brendan M Duggan, Jehad Almaliti, Pierre-Marie Allard, Prasad Phapale, Louis-Felix Nothias, Theodore Alexandrov, Marc Litaudon, Jean-Luc

Wolfender, Jennifer E Kyle, Thomas O Metz, Tyler Peryea, Dac-Trung Nguyen, Danielle VanLeer, Paul Shinn, Ajit Jadhav, Rolf Müller, Katrina M Waters, Wenyuan Shi, Xueting Liu, Lixin Zhang, Rob Knight, Paul R Jensen, Bernhard O Palsson, Kit Pogliano, Roger G Linington, Marcelino Gutiérrez, Norberto P Lopes, William H Gerwick, Bradley S Moore, Pieter C Dorrestein, and Nuno Bandeira. "Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking." In: *Nature biotechnology* 34.8 (2016), pp. 828–837. DOI: 10.1038/nbt.3597. URL: http://www.ncbi.nlm.nih.gov/pubmed/27504778.

[70] Darren Kessner, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. "ProteoWizard: open source software for rapid proteomics tools development." In: *Bioinformatics* 24.21 (Nov. 2008), pp. 2534–6. DOI: 10.1093/bioinformatics/btn323. URL: http://www.ncbi.nlm.nih.gov/pubmed/18606607.

[71] Wout Bittremieux, Christopher Chen, Pieter C. Dorrestein, Emma L. Schymanski, Tobias Schulze, Steffen Neumann, Rene Meier, Simon Rogers, and Mingxun Wang. *Universal MS/MS Visualization and Retrieval with the Metabolomics Spectrum Resolver Web Service*. preprint. Bioinformatics, May 10, 2020. DOI: 10.1101/2020.05.09.086066. URL: http://biorxiv.org/lookup/doi/10.1101/2020.05.09.086066 (visited on 09/29/2022).