Multi-reference spectral library yields almost complete coverage of heterogeneous LC-MS/MS data sets

Constantin Ammar[1,2], Evi Berchtold[1], Gergely Csaba[1], Andreas Schmidt[3], Axel Imhof[2,3], Ralf Zimmer[1,2]

[1] Institute of Bioinformatics, Department of Informatics, Ludwig-Maximilians-Universität München, Amalienstrasse 17, 80333 München, Germany

[2] Graduate School of Quantitative Biosciences, Ludwig-Maximilians-Universität München, Feodor-Lynen-Str. 25, 81337 München, Germany

[3] Zentrallabor für Proteinanalytik (Protein Analysis Unit), Ludwig-Maximilians-Universität München, Groβhaderner Straβe 9, 82152 Planegg-Martinsried, Germany

Correspondence should be addressed to Ralf Zimmer.

Email: Ralf.Zimmer@bio.ifi.lmu.de

Phone: +49 89 2180 4052

Fax: +49 89 2180 99 4052

Running Title: Multi-reference spectral libraries

Abbreviations

DIA: data independent acquisition

DDA: data dependent acquisition

SRM: selected reaction monitoring

MRM: multiple reaction monitoring

PRM: parallel reaction monitoring

SWATH: sequential window acquisition of all theoretical mass spectra

CIP: characteristic intensity pattern

MCIP: multiple characteristic intensity patterns

FA: formic acid

TFA: trifluoroacetic acid

ACN: aceonitrile

PSM: peptide spectrum match

cps: counts per second

NRFV: normalized replicate fragmentation vector

TPP: trans proteomic pipeline

SVM: support vector machine

Summary

Spectral libraries play a central role in the analysis of data independent acquisition (DIA) proteomics experiments. DIA experiments require spectral libraries, as most current methods cannot apply traditional peptide identification via database searching on DIA data. A central assumption in current spectral library tools is that a single characteristic intensity pattern (CIP) suffices to describe the fragmentation of an unmodified peptide in a particular parent ion charge state (peptide charge pair). However, we find that this is often not the case.

We analyze a heterogeneous dataset of 440.000 MaxQuant - preprocessed peptide spectra from a QToF mass spectrometer, stemming from over 100 different LC-MS/MS runs. The dataset corresponds to 10.580 peptide charge pairs, which have each been measured and identified at least 20 times. We demonstrate that the same charged and unmodified peptide can fragment in multiple reproducible ways, even within the same LC-MS/MS run. We integrate multiple reference CIPs (MCIPs) in our model library and observe a >99% coverage of replicate fragmentation spectra for 95% of peptide charge pairs (using up to four CIPs). Using a single CIP (as in current spectral library approaches), we find >99% coverage for only 50% of the peptide charge pairs.

Our approach achieves substantially greater sensitivity in comparison to the popular SpectraST library generation tool. Using randomized decoy spectra, we demonstrate that identification accuracy of the MCIP approach is improved by up to 12% compared to a single CIP approach. We test the MCIP approach on a SWATH data set and observe a ~30% increase in peptide recognition. We conclude that including MCIPs in spectral library approaches would yield increased sensitivity without compromising the false discovery rate.

Introduction

For the standard *data dependent acquisition* (DDA) approaches of proteomics, selected *precursor ions* are isolated in a small mass window and subsequently submitted for fragmentation and MS2 measurement (1,2). In the most widely used DDA approach (Top-N), also called *shotgun proteomics*, the fragmentation spectra of the N most intense precursor ions are acquired in each duty cycle and analyzed by scoring the *mass to charge* (m/z) values of the most intense fragment peaks against a theoretical prediction of m/z values of fragment ions derived from sequence databases (3). The theoretical m/z values of fragment ions are discriminative as, in most cases, each peak in the MS2 fragmentation spectrum stems from the same precursor ion. Additionally, the m/z value of the submitted precursor ion is known, which narrows down the number of possible matches in the sequence database. This high confidence approach comes at the price of decreased coverage, as all peptide precursors not selected for fragmentation cannot be analyzed (4). As precursor ion selection can be described as semi-random (5), DDA approaches are also problematic for quantification, as a peptide measured in a first run, might not be selected in a second run, even though it is abundant.

Selected reaction monitoring (SRM, alternatively multiple reaction monitoring (MRM) or parallel reaction monitoring (PRM)) approaches (6, 7) address this problem by a fixed preselection of precursor ions. This approach allows very sensitive and accurate quantitation of proteins, however, due to preselection of features, the overall coverage of the proteome is low and all information that is excluded from the feature list is lost.

Data independent acquisition (DIA) approaches try to overcome these limitations by leaving out the preselection of precursor ions (8-10). To reduce complexity, many implementations scan repeatedly over a wide m/z range in short fragmentation windows (11-17). The whole m/z range is covered in a few seconds cycle time. Due to their high sensitivity, DIA approaches like the *sequential window acquisition of all theoretical mass spectra* (SWATH) (16, 17) have recently gained attention.

In general, the possibilities for spectral searches via sequence databases are challenging for DIA data (18, 19) due to the ambiguity of m/z values in complex peptide mixtures. Thus, many commonly used approaches rely on *spectral libraries,* also considering fragment ion intensities (17). These libraries are obtained from DDA proteomics experiments, by generating a *characteristic intensity pattern* (CIP) of (m/z – ion intensity) pairs from confidently identified MS2 spectra for each peptide in a distinct charge state (peptide charge pair) (20 - 23).

A library pattern must be constructed such that it is sufficiently specific (few false positives) while maintaining high sensitivity (few false negatives). A library of CIPs is then compared to the measured fragmentation spectrum using a *similarity measure*. Most current approaches for the construction of library patterns employ the scoring measure *dot product* (24 – 27) or the related *spectral contrast angle* (28 – 30). To our knowledge, all tools try to approximate one unique CIP from the available measured fragmentation spectra.

Prior to their use in DIA approaches, spectral libraries have been employed to speed up and increase confidence in peptide recognition (20) and currently, large spectral repositories (31-34) exist. In current DIA applications like OpenSWATH (17), *chromatography-based* scores are used to identify consistent MS2 fragmentation spectra, which are then matched with a library CIP. Hence, having an accurate spectral library and a highly reproducible and calibrated LC system are key factors determining the quality of a DIA experiment. Nevertheless, even under optimal experimental conditions, the applied collision energy differs between the acquisition modes. In DDA mode, the charge of the precursor ion is determined during a MS1 survey scan before submission to fragmentation. The collision energy is adjusted to an average optimum depending on mass and charge using machine-specific equations for each charge. In SWATH-acquisition, the collision energy is determined using the same linear equation under the assumption of a default charge state 2+ (16), resulting in sub-optimal fragmentation energies for peptides with higher charges. In MRM, fragmentation energies are often optimized for each precursor – fragment pair (*transition*) in order to maximize the signal intensity, in PRM experiments, a single fragmentation with optimized collision energy is performed and all fragments are acquired in one spectrum.

In the context of these developments, improving spectral libraries gains renewed importance. In this study, we present a systematic analysis of fragmentation spectra identified with high confidence, by generating and evaluating a model spectral library. Our *Multiple Characteristic Intensity Pattern (MCIP)* method is similar to the SpectraST approach by Lam et al. (20), but differs in three key points: (i) We conduct our library generation on MaxQuant (35) preprocessed peptide identifications without modifications and considering only b- and y- ions (with molecular losses). This reductionist approach enables us to focus on the core properties of peptide fragmentation, while utilizing the preprocessing algorithms of the MaxQuant software, which have shown to drastically improve peptide recognition (35). SpectraST uses semi-raw (.mzXML) fragmentation spectra for the generation of spectral libraries, without further preprocessing (23). (ii) As we have already preprocessed the spectra, we can apply either a ranking by signal to noise previous to the clustering or use an unranked approach. In both cases, we apply a systematic clustering until all spectra are contained in a cluster and retain all clusters involved. (iii) We determine one CIP from each cluster. This can yield more than one CIP per peptide charge pair.

Table 1 gives an overview and more details of our MCIP method and compares it to state-of-the-art approaches.

[Table 1]

In the following analysis, we use peptides, which have been measured and identified at least 20 times across several experiments, yielding ≥ 20 *replicate* fragmentation spectra for each peptide charge pair. To our knowledge, there is currently no spectral library analysis, which restricts to such a high number of replicate spectra. We observe that for many peptide charge pairs, a limited set of distinct patterns, which are each highly reproducible, is measured within and across different LC-MS/MS runs. In this paper, we describe the MCIP method in detail, evaluate its performance compared to a single CIP approach (as is currently the default for spectral library approaches) and show that MCIP yields significantly increased sensitivity with similar specificity in benchmarks. We demonstrate that the increase in spectral recognition by MCIP also applies to SWATH data. Finally, we investigate possible experimental factors influencing the MCIP phenomenon (different reproducible fragmentation of peptide charge pairs and the measurement of MCIPs). We carry out tailored LC-MS/MS runs, where we vary the collision energy and width of the precursor isolation window. These runs show an enrichment of differently fragmenting spectra for wider isolation windows, suggesting that interferences with the background matrix influence the reproducibility of peptide fragmentation. A software to check MaxQuant processed proteomics runs is available on https://www.bio.ifi.lmu.de/forschung/proteomics/mcip/index.html.

Experimental Procedures

Proteomic analysis via LC-MS/MS on Q-TOF mass spectrometer

Samples were injected into an Ultimate 3000 HPLC system (Thermo Fisher Scientific). For nano-reversed phase separation of tryptic peptide mixtures before MS analysis, peptides were desalted on a trapping column (5 x 0.3 mm inner diameter; packed with C18 PepMap100, 5 μm particle size, 100 Å pore diameter, Thermo-Fisher Scientific). The loading pump flow of 0.1 % formic acid (FA) was set to 25 μl/minute with a washing time of 10 min under isocratic conditions. Samples were separated on an analytical column (150 x 0.075 mm inner diameter; packed with C18RP Reposil-Pur AQ, 2.4 μm particle size, 100 Å pore diameter, Dr. Maisch) using a linear gradient from 4 % to 40 % B in 170 min with a gradient flow of 270 nl/minute.

Solvents for sample separation were A 0.1 % FA in water and B: 80 % acetonitrile (ACN), 0.1 % FA in water. The HPLC was directly coupled to the 6600 TOF mass spectrometer using a nano-ESI source (both Sciex). A data-dependent method was selected for MS detection and fragmentation of eluting peptides comprising one survey scan for 225 ms from 300 to 1800 m/z and up to 40 tandem MS scans for putative precursors (100-1800 m/z). Precursors were selected according to their intensity. Previously fragmented precursors were excluded from reanalysis for a timespan between 10 and 50 seconds, depending on the experiment (see supplemental table 2). Rolling collision energy setting was enabled allowing to perform fragmentation at the optimized collision energy for the peptide charge pairs. Precursor charge states from +2 to +5 were specifically detected.

SWATH data acquisition

Peptides from tryptic digestion were resuspended in 10 µl 0.1 % trifluoroacetic acid (TFA) and injected into a Ultimate 3000 nano-chromatography system equipped with trapping column (C18 AcclaimPepMap, 5 x 0.2 mm, 5 µm 100 Å) and a separation column (C18RP Reposil-Pur AQ, 150 x 0.075 mm x 2.4 µm, 100 Å, Dr. Maisch, Germany) poured into a nano-ESI emitter tip (New Objective, Woburn MA). After washing for 10 min on the precolumn with 0.05 % TFA, peptides were separated by a linear gradient from 4 % to 40 % B (solvent A 0.1 % FA in water, solvent B 80 % ACN, 0.1 % FA in water) for 150 min at a flow rate of 270 nl/min. Eluting peptides were detected on a Triple TOF quadrupol-TOF hybrid mass spectrometer (Sciex, Framingham, MA). First, a mixture of all conditions was run in data-dependent mode to generate an ion library for the data-independent SWATH measurements and optimize the isolation window distribution over the mass range for SWATH-data acquisition. Data-dependent acquisition consisted of a survey scan and up to 40 tandem MS scans for precursors with charge 2-5 and more than 200 counts per second (cps) abundance. Rolling collision energy was set to generate peptide fragments. The overall cycle time for the DDA experiment was 2.676

seconds. Previously analyzed precursors were excluded from repeated fragmentation for 30 seconds employing a mass window of 20 ppm around the precursor mass.

MS data with data-independent SWATH acquisition were generated using the same HPLC conditions as used for the generation of the ion library. Based on the distribution of the m/z values of identified peptides in the ion library, the mass range from 300-1200 m/z was split into 40 SWATH mass windows. First, precursors were monitored from 300-1500 m/z in a survey scan of 50 ms, followed by the SWATH data acquisition for 65 ms/mass window, resulting in an overall cycle time of 2.7 seconds. The fragmentation energy was adjusted to fragment 2+ charged ions in the center of the mass window and a collision energy spread over 7 units was allowed.

Targeted LC-MS/MS experiments to study fragmentation settings

To investigate the influence of isolation width and collision energy on occurrence of distinct fragmentation patterns, targeted LC-MS/MS runs were carried out on a subset of 30 peptides associated with multiple fragmentation patterns and 30 control peptides. Control peptides were selected with similar number of identified mass spectra, to avoid comparison to highly abundant peptides. For this experiment, a special SWATH setup with a 60 min gradient on a uLC system was employed. The SWATH windows were specifically set for each precursor to meet the requirements of the study. Fragmentation settings were kept stable throughout a single run. 5 different collision energy settings and 5 different isolation windows were selected and measured in two replicates per condition. The optimal collision energy depends on correct charge state assignment, to study if too high or low collision energy alters the peptide fragmentation, we applied collision energies -3 V and -1.5 V lower than the calculated optimal collision energy as well as 1.5 V and 3 V higher. As reference the calculated optimal collision energy was selected, for this study the isolation window was kept at 1 Da around the precursor. An sub-optimally designed isolation window can cause co-isolation of other peptides and therefore mixed fragment ion spectra. Therefore, we applied isolation windows of 1 to 5 Da to study the effect of the matrix background on reproducibility of fragmentation. Peptide sequences and mass spectrometer settings are listed in supplemental table T5.

Data analysis of data-dependent LC-MS/MS experiments

Data-dependent experiments performed on the QToF mass spectrometer were analyzed in MaxQuant (version 1.5.1.2 and higher) using the Andromeda search engine (36) and a sample specific protein database in FASTA format (see supplemental table 1). The settings for database search were as follows: fixed modification carbamidomethylation of cysteine, variable modifications oxidation of methionine and acetylation at the protein N-terminus ; Δmass = 30 ppm for precursors for the first search and 6 ppm for the second search, Δmass = 60 ppm for TOF fragment ions, enzyme trypsin with specific cleavage and max. two missed cleaveages. Peptide hits required a minimum length of 7 amino acids and a minimum score of 10 for unmodified and 40 for modified peptides. PSM-FDR was set to 1%.

MaxQuant preprocessing included mass-centroiding of peaks and corresponding intensity adaption, de-isotoping and detection of co-fragmented peptides (35). The results were returned as msms.txt files, containing the relevant spectral information of fragment ion intensities, retention times, fragment masses as well as charge and modification states of the identified peptide.

The MS proteomics data, including MaxQuant results have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the data set identifiers PXD005060, PXD005063, PXD005100, PXD005111, PXD006245 and PXD006691. More details on the datasets used are displayed in supplemental table 1. Supplemental table 2 contains all information on the peptide identifications.

The spectra have been uploaded to the MS-Viewer database and the search keys are listed in supplemental table T3.

Selection of processed fragmentation spectra

Peptides were separated by charge into *peptide charge pairs*, because differences in the charge state significantly alter the fragmentation pattern (see supplemental Fig. S1). Only peptide charge pairs, which had at least 20 replicate spectra (see supplemental Fig. S2), were included to enable the statistical analysis of repeated fragmentation of chemically identical peptides. In our main analysis, we restricted our MCIP approach to only b- and y- ions in all charge states with different molecular losses (examples: b3, y4-NH_3, y6(2+), b5(2+)-H_2O). In additional supplemental analyses, we also consider other ions for comparison (supplemental Fig. S6). Modified peptides were excluded.

Import of semi-raw fragmentation spectra

To quantify the impact of using all peaks without filtering, an additional analysis with raw spectra was carried out. To assess the influence of the preprocessing method, two different methods of preprocessing the data were applied. In the first approach, the raw spectra were imported from the MaxQuant .apl files contained in the “andromeda” folder in the MaxQuant output folder. The files were parsed using a self written parser. From these files we extracted a list of m/z values with corresponding intensities, without b- and y- ion annotation for each spectrum. The spectra were assigned to their respective MaxQuant identification via the spectrum index. In the second approach, the raw .wiff files were processed into the .mzXML format with the MSConvert tool (38) without any additional filters (yielding profile data) and analogously imported using a self written parser and assigned to their respective MaxQuant identification via the spectrum index. To avoid problems with false indexing, the .mzXML files were additionally scored with the in-house scoring ReScore described below. The differences between the preprocessing methods can be seen in supplemental figure S6.

Assessment of the similarity of fragmentation spectra

The similarity among spectra of the same peptide charge pair (*replicate* spectra) can be used as a measure to characterize the fragmentation behavior of peptide charge pairs. As spectra are vectors of (m/z, intensity)-pairs, they can differ in the m/z-values (different peaks) or their intensities, or both. If replicate spectra are highly similar, the fragmentation process can be described as distinct and reproducible. If replicate spectra are very dissimilar, a reason could be different fragmentations of the precursor ion.

To assess similarity between replicate spectra, all replicate spectra (at least 20, see previous section) available for a peptide charge pair were pairwise compared to each other.

Each fragmentation spectrum was represented as a *normalized replicate fragmentation vector* (NRFV) *I* of length 1

[equation 1]

with i_1 to i_n denoting the intensities in the pattern and the indices of the vector implicitly denoting the different fragmentation ions (m/z values). To get vectors of equal length, each fragmentation ion with intensity >0 in any of the replicate spectra was included in every vector. Imputed 0 values were used, if a corresponding ion was not observed. A best bipartite matching method was used for raw spectra at different ppm precisions. Only vectors with length >4 were used. Each vector was normalized to length 1 (unit vector).

After determining which intensities were included in the NRFVs, the *spectral similarities* between all NRFVs of a peptide charge pair were assessed in a pairwise fashion. For each pair *X* and *Y* of NRFVs, the *dot score* was calculated using the dot product *similarity measure DP*, defined as

[equation 2]

with xk and yk denoting the k-th element of X and Y, respectively. If the dot score was calculated between b- and y- ions of the spectra, it was called dot score, if it was calculated among semi-raw spectra it was called raw dot score. For the raw dot scores, the score was determined for four different ppm precisions (50 ppm, 100 ppm, 150 ppm, 200 ppm).

A pair of fragmentation spectra was called similar, if the dot score of their two corresponding NRFVs was larger than a predefined *similarity threshold* (see below).

Centroid clustering and CIPs

A central goal of this study is to find a minimal set of *characteristic intensity patterns* (CIPs), able to characterize all observed fragmentation spectra of a peptide charge pair. In order to derive these, a *centroid clustering* approach was employed to determine *clusters* of similar NRFVs. For each NRFV, the *neighborhood* (all fragmentation spectra with a similarity score greater than the chosen similarity threshold) was determined. The NRFV corresponding to the spectrum with the best signal to noise ratio (defined via the average intensity of the 2nd to 6th highest peak divided through the median of the remaining peaks) was defined as a CIP, analogous to the SpectraST approach (20). Additionally, also NRFVs with the largest number of neighbors were defined as CIPs (see supplemental figure S5). If not all NRFVs were neighbors to this CIP, it was removed with all its neighbors and the procedure was repeated on the remaining NRFVs. This determined more CIPs for the peptide charge pair in question. Depending on the number of CIPs resulting from this procedure, each peptide charge pair was assigned either a *single CIP* (all spectra of a peptide charge pair assembled in a single cluster) or *multiple CIPs* (MCIPs). The CIPs were referred to by size of their respective cluster: CIP_1 corresponds to the largest cluster, CIP_i to the i-th largest cluster.

Spectral coverage

The *spectral coverage* was introduced as a measure for the sensitivity of the approach. A spectral library was constructed with the entries for each peptide charge pair consisting either of a single CIP of the largest cluster, or of MCIPs $\{CIP_1, CIP_2, ..., CIP_n\}$ of the n largest clusters. The single CIP or each element of the MCIPs $\{CIP_1, CIP_2, ..., CIP_n\}$ was then compared to all NRFVs of the peptide charge pair using the dot score. If the dot score was above the similarity threshold for any of the CIPs, the respective spectrum was marked as covered. The spectral coverage denotes the fraction of replicate spectra covered.

Comparison with SpectraST

A comparison of the spectral coverage with the popular SpectraST search engine (23) was carried out. For this, raw spectra (.wiff format) of human cell lines (see supplemental table 1) were submitted to the Trans-Proteomic Pipeline (*TPP*) (37) software suite v.4.8.0, which contained all tools used in the further processing. In parallel, the same .wiff files were processed using the MCIP approach described above. The .wiff files submitted to the TPP were converted into the .mzXML format using msconvert (38) and files were searched using the spectrum search engine X!Tandem (39) under the “isb_default_input_kscore.xml” configuration included in the TPP. The Ensembl v.75 *Homo sapiens* GRCh37 sequence database was used. The resulting .pep.XML output files were evaluated using PeptideProphet (40). The PeptideProphet ranked .pep.XML output files were submitted to SpectraST in library create mode using the default configurations. The resulting raw library was processed to a consensus library using the corresponding option included in SpectraST. The consensus library mode was chosen, because it has been shown to give the highest number of positive identifications (20). The consensus library was then quality filtered using the highest quality level (option -cL5) in SpectraST.

In order to systematically compare SpectraST and the MCIP approach, a similar spectral coverage approach as discussed above was used for SpectraST. The PeptideProphet ranked .pep.XML file was parsed and the input spectra for each SpectraST library entry were determined. Then, the SpectraST library was used to search these spectra in the .mzXML files.

The search is applied to the same data used for the construction of the library. This idealized search scenario conservatively measures whether the CIP constructed by SpectraST suffices to recapture all previously identified replicate spectra of a peptide. The resulting .xml output file of the search was parsed and, analogous to the MCIP library creation approach, a spectrum was marked as covered, if the dot score of the library consensus pattern with the spectrum was greater than 0.6.

To compare with the MCIP approach, only spectra that were identified in both the MaxQuant and the PeptideProphet output were further processed. Additionally, only peptides that had at least 10 replicate spectra in both the MaxQuant and PeptideProphet output were included. The number of replicate spectra was lowered to 10 (from 20) to increase the number of identified peptide charge pairs.

Benchmarking via cross validation

To conduct performance testing, the setup visualized in Fig. 1 was used to determine false positives and false negatives. The replicate spectra of each peptide charge pair were split into two fractions. The first fraction consisted of 20% of the spectra and each spectrum was assigned a decoy spectrum $\vec{P}_{\mathrm{decoy}}$ which contained m/z-shuffled intensities of the original spectrum. By shuffling the spectra, the total intensity and the m/z values were preserved, while the spectrum represented changed completely. A 1:1 mixed test set containing original and decoy spectra was then generated. The second fraction consisted of the remaining 80% of spectra. On this fraction, CIP(s) were created as described in the previous sections. The CIP(s) were then similarity scored against the test set using the dot score. A similarity score below the similarity threshold for an original spectrum $\vec{P}_{\mathrm{orig}}$ was marked as false negative, a score above the threshold with a decoy spectrum $\vec{P}_{\mathrm{decoy}}$ was marked as a false positive. The m/z-shuffling approach is similar to the method employed by Lam et al. (41), where counting of decoy matches is used library wide to estimate the FDR. Different from Lam et al. (41), each set of replicate spectra was individually checked via fivefold cross validation in this study. This allowed estimating the relative fractions of false positives and false negatives per peptide charge pair, rather than library wide.

[Fig. 1]

Validation of the global similarity threshold

A *global similarity threshold* of dot score 0.6 was adapted from the SpectraST search engine (23) and was subsequently validated using the sampling analysis approach discussed above. This was done to check, whether this threshold would give overall discriminative results. Each spectrum in the dataset was represented as a NRFV and assigned 1000 differently shuffled decoy vectors. Each NRFV was then dot scored against each decoy vector, which resulted in a distribution of 1000 *shuffled dot scores* for each NRFV. From each distribution of shuffled dot scores, a *local discriminative dot score* was extracted, such that less than 5% of the shuffled dot scores were above this threshold (in other words, the 95% quantile was extracted). This meant, that the use of this dot score would result in 5% acceptance of decoy spectra for a particular NRFV. All local discriminative dot scores were collected, which resulted in a new distribution. From this new distribution, again the 95% quantile was extracted (see supplemental Fig S3). This 95% quantile was 0.62 in this study, which agreed well with the global similarity threshold of 0.6. The approach of extracting two quantiles was taken, because the distribution of shuffled dot scores varied distinctly for different spectra. Hence, taking only one quantile on the distribution of all shuffled dot scores of all spectra combined would result in some spectra (the spectra with generally large shuffled dot scores) being completely ambiguous.

Testing the implications for SWATH data

To get comparable results with SWATH, a whole cell extract from the fruitfly *Drosophila simulans* that had been analyzed using both DDA and SWATH was used (see supplemental table 1, Id 5). The input samples for both measurements were identical, hence the peptides discovered in the DDA run could be expected to be present in the SWATH dataset at the same retention time (+/- 1min). The peptides identified in the *Drosophila simulans* DDA run were compared against all other DDA results and only peptides with at least 20 replicate spectra observed in the other DDA results were included in the analysis. This applied to 29% of all identified peptides in the *Drosophila simulans* DDA run. The CIPs derived from the other DDA results were then separately dot scored against the output spectra of the DDA and the SWATH run in their respective retention time windows (similarity threshold 0.6).

Processing of targeted LC-MS/MS runs for CE and isolation window study

Due to the special data acquisition setup described above, the output of the targeted LC-MS/MS experiments was not accessible to standard DDA processing via MaxQuant. The .wiff files were converted to .mzXML using MSConvert (38) and the .mzXML files were then processed using an in-house scoring method, termed *ReScore*. ReScore is based on the scoring described in the publication of the MaxQuant search engine Andromeda (36). Using IDA runs that were carried out along with the the targeted LC-MS/MS runs on the same standardized HeLa Pierce lysate (PXD006691), the scores were compared with Andromeda. The scores show strong correlation with the Andromeda scoring and the vast majority of Andromeda scores is higher than the corresponding ReScore (supplemental Figure S7). Hence, a certain ReScore cutoff can be used as a reliable cutoff for the Andromeda score.

Results

Library generation settings

The spectral library was solely generated from DDA data only, with the explicit runs marked in supplemental table 1. For the instrument, the standard configurations as recommended by Sciex were applied to all setups with the vast majority of parameters fixed between all runs. Different settings were only applied to the parameters “Exclude for:” (range 10s - 50s), “Mass tolerance:” (15ppm – 50ppm), “Switch After” (30 spectra – 40 spectra) and “With intensity greater than” (100 - 150). Rolling collision energy was set in all cases. The specific parameters for each input sample are listed in supplemental table T2.

MCIPs are frequent and reproducible

We used the centroid clustering approach on all peptide charge pairs with 20 or more replicate spectra. The centroid(s) of the neighborhood(s) of normalized replicate fragmentation vectors (NRFVs) were taken as the characteristic intensity pattern(s) (CIP(s)) for each peptide charge pair. For reasonable similarity thresholds, we observed a large fraction of peptide charge pairs, which had multiple CIPs (MCIPs). Additionally, we applied a *rematch constraint*, meaning that a minimum number of spectra (up to six) should match with CIP_2 (corresponding to the second largest cluster) (Fig. 2a). We called these peptide charge pairs *MCIP peptide charge pairs*. For our global similarity threshold of 0.6 and using a rematch constraint of at least four replicate spectra, we observe 23% MCIP peptide charge pairs. Even at an over-sensitive, low specificity similarity threshold of 0.4 (see supplemental Fig. S3), we observe MCIP peptide charge pairs in approximately 9% of all peptide charge pairs. At very high similarity thresholds, over-clustering occurs, with the number of spectra per cluster decreasing. A decrease in the percentage of MCIP peptide charge pairs then becomes visible, because the rematch constraint is not fulfilled anymore.

We further tested if the MaxQuant false discovery rate (FDR) of 1% could significantly influence the phenomenon. This was necessary, because false identifications by MaxQuant could cluster separately from the correctly identified spectra and give incorrect CIPs. As the MaxQuant FDR is defined per LC-MS/MS experiment, we calculated the fraction of spectra outside of the largest cluster 1 for every experiment (supplemental Fig. S4), which shows, that this fraction is often more than 30 times the MaxQuant FDR at a threshold of 0.6. Additionally, it is shown that even with a rematch constraint of four replicate spectra, the MaxQuant FDR is often exceeded more than tenfold.

As an additional test, we checked whether different fragmentation patterns also occurred within the same experiment, considering only replicate spectra with precursor m/z difference <0.01 and a retention time difference of <2min. Additionally we applied a rematch constraint of four replicate spectra. Even with these strong constraints, we found 712 examples of peptides with diverging patterns within the same experimental run. Examples of diverging spectra within the same run can be interactively viewed and compared via the viewer on the webpage https://www.bio.ifi.lmu.de/forschung/proteomics/mcip/index.html.

As a visualization of MCIPs, NRFVs of cluster 1 (blue) and cluster 2 (red) are plotted on top of each other in Fig. 2b. The examples show that clusters of replicate fragmentation spectra can be very different from each other, while being highly similar within the cluster. An interactive viewer on differing clusters is also available on the webpage. To further understand the clustering observed in these figures, we searched for possible macroscopic parameters correlating with the different clusters. However the machine settings were fixed over the different experiments in this case (see supplemental table T2).

We examined the dependence of the number of CIPs on the number of replicate patterns. This revealed that the number of CIPs per peptide charge pair saturates quickly and is rarely larger than six, even for peptide charge pairs for which hundreds of replicate patterns are available (Fig. 1c). The average number of CIPs is about 2 (1.96).

[Fig. 2]

Usage of MCIPs yields almost complete spectral coverage

To test the sensitivity of the MCIP approach, we assessed the spectral coverage (see method section) for all peptide charge pairs. The spectra were either compared with one CIP (in accordance with the current library approaches) or with MCIPs. Fig. 3a shows a striking improvement upon successive integration of more and more CIPs (red up to orange line) until near-complete coverage is reached. The largest gain is visible upon integration of the second CIP (blue line), which corresponds to the second largest cluster. This improvement clearly shows that significantly higher peptide recognition is possible by simply integrating one more CIP into the library. Improvement is seen for adding additional CIPs, but is decreasing with each additional cluster. This effect is for one an intrinsic property of choosing the next CIP corresponding to the next smaller cluster. Nevertheless, the strong variation, especially between the first and the second cluster (red and blue line respectively) is highly unlikely due to random divergences. The cumulative distribution of the relative fractions of the different clusters is displayed in Fig. 3b.

[Fig. 3]

Direct comparison with SpectraST shows significantly increased spectral coverage

We compared the spectral coverage of our MCIP approach with the coverage of the SpectraST (20) spectral search engine, which is among the most popular in the field (42).

As is described in the methods section, a setup similar to the spectral coverage estimation was introduced for SpectraST. A SpectraST library was constructed on three human data sets (see supplemental table 1). In order to make the identification as simple as possible for SpectraST and thereby the coverage as high a possible, we tried to re-identify the spectra used for the construction of the SpectraST library on the very same human data sets. This allowed us to assess the spectral coverage. We only considered spectra, which had been used for library construction in our approach as well as in SpectraST. We further constrained this set of common spectra by only including peptide charge pairs with at least 10 replicate spectra. This resulted in a total of 402 peptide charge pairs. We then assessed the spectral coverage of both methods on this data set.

Fig. 4 shows a significantly improved performance of the MCIP approach (red to green line) in comparison to SpectraST (dark blue). The baseline identification is lower for the SpectraST approach. When comparing the performance of CIP_1 (red line) with SpectraST, we first see a higher performance of CIP_1 and an intersection of the spectral coverage of both methods at around 83% coverage of replicate spectra. This means, that the same number of peptides has 83% of its replicates covered with the SpectraST consensus spectrum and with CIP_1. Above 83%, SpectraST performs slightly better than CIP_1. When integrating CIP_2 (light blue line), SpectraST is significantly outperformed, as almost all peptides are fully covered with the two CIPs. The third CIP still gives a slight increase in coverage. The partially lower performance of CIP_1 in comparison with SpectraST might be explained by considering, that due to the systematic clustering, the spectra corresponding to the other clusters are filtered out and this information is hence missing when only using one CIP. When integrating the information about the other clusters, the strength of the MCIP approach becomes visible.

[Fig. 4]

MCIPs increase sensitivity without affecting specificity

As has been shown in the previous sections, MCIPs are needed to describe all replicate spectra for many peptide charge pairs. Obviously, MCIPs improve sensitivity. Nevertheless, they may lead to reduced specificity. Hence, we employed an accuracy test by first generating CIPs on 80% of the replicate spectra and then scoring these CIPs against a mixture of the remaining replicate spectra and shuffled decoy spectra. This allowed to distinguish true positives (match of CIP with replicate spectrum) from false positives (match of CIP with decoy spectrum). The procedure is described in more detail in the methods section

Fig. 5a) shows that the overall accuracy for MCIPs (blue and green line) increases significantly in comparison to a single CIP (red line). For >99% of peptide charge pairs, the minimum accuracy increases by 12% when integrating all CIPs available for each peptide charge pair in the spectral library (green line). In Fig. 5b), we see (as expected from the spectral coverage results) a strong decrease in false negatives upon integration of MCIPs. At the same time, we see that the false positive rate displayed in Fig. 5c) is almost not affected at all.

[Fig 5]

Implications for SWATH data

One of the current applications of spectral libraries is the analysis of SWATH data (17). In this setup, CIPs are matched with more complex MS2 spectra. We assessed, whether MCIPs can improve the identification rate as compared to using only one CIP.

As described in the methods section, we used a dataset, where the same sample had been identified using DDA and SWATH. We then utilized this setup to derive a spectral library of peptides, which was expected to be in the SWATH data set.

We then searched the library patterns against the DDA run as well as the SWATH run, with the fraction of non-identified spectra ("errors") plotted against the similarity threshold in Fig. 6. For the DDA run, the results are analogous to the results already presented, with a significant improvement of identification upon integration of MCIPs (red and blue line, respectively).

For the SWATH run, we observe lower baseline identification, with approximately 20% of the patterns not being identified at all. This might be due to the higher noise in the SWATH patterns, likely caused by smaller individual fragmentation times as well as cross-fragmentation of different peptides. Nevertheless, also for the SWATH data set, we observe very similar effects when comparing the single pattern approach (green) with the MCIP approach (magenta), with an ~30% increase in identification accuracy at reasonable similarity scores (e.g. at 0.6: 29%).

[Fig. 6]

Discussion

By the analysis of peptide charge pairs with many (≥20) replicate spectra, we could show that multiple CIPs (MCIPs) are observed for almost 25% of identified peptides in our data and occur even within the same LC-MS/MS run.

Even though the focus of this study lies on the quantification and computational solution of this frequent data acquisition problem, we also examined a variety of factors possibly being responsible for the phenomenon. Especially, different CIPs within the same run are unexpected and a wide range of possible explanations exists. One intuitive explanation is that the differently fragmenting peptides are assigned the wrong charge state in the MS survey scan in DDA mode. Due to the wrong charge state assignment, the collision energy applied to fragment the peptide is too high or low and it is therefore over- or underfragmented, respectivley. Precursor charge state and collision energy were not stored in the obtained raw files, hence we were not able to directly check this. However, we carried out targeted LC-MS/MS runs, where we specifically adjusted the applied collision energy. Our results show that dot scores are surprisingly robust over a range of -3V to +3V, which covers differences in collision energy settings caused by wrong precursor charge state assignment and only slight effects are seen on the occurrence of MCIPs (supplemental figures S10 and S11). Further, we examined dependence of fragmentation stability on a variety of parameters likely to be relevant for peptide fragmentation by analyzing new DDA runs which were carried out on the same standardized lysate as the targeted runs. We identified significant shifts depending on parameters correlating with precursor isolation purity and intensity (supplemental figures S8 and S9). In general, high fragmentation stability (dot score with CIP1 >0.9) correlates with high peak purity and intensity, while a clear distinction between medium stability (dot score with 0.6<= CIP1 <=0.9) and differently clustering spectra (dot score with CIP1 <0.6) is not possible.

The observed behavior suggests that there is a connection between the fragmented peptide and the *background matrix (coeluting peptides in the same isolation window*), which influences the fragmentation behavior. For high abundant peptides, this interference has only little effect. For low abundant peptides however, it has already been shown that ion interferences with the background matrix can alter the fragmentation spectrum (45) and simulations on the ion interferences suggest an abundance of around 10% of interfering ions for small isolation windows (16), which is in a similar range as the number of MCIPs we observe. This would also explain the stochastic nature as well as the reproducibility of the phenomenon, as different fragmentation would only occur, if a particular interfering ion for a MCIP peptide charge pair would co-elute in the LC-column. This co-elution is to some extent reproducible, but critically depends on the background matrix, as well as on the peak overlaps between MCIP and interfering peptide. Also, as not many interfering ions are expected to exist for a given MCIP peptide, the limited number of clusters observed would be explained.

We carried out additional targeted LC-MS/MS runs to investigate the influence of the background matrix, therefore the precursor isolation width was varied between 1Da and 5 Da. For broader isolation windows, we observed a systematic enrichment in differently fragmenting spectra, while, as expected, the phenomenon is still seemingly stochastic (supplemental Figures 10 and 11).

To ensure reproducibility of the patterns, we only classified a peptide charge pair as an MCIP peptide charge pair, if the second cluster had been matched at least four times (Fig. 2a). Four spectra should be a sufficient number for the creation of a library entry, as common library generation tools usually have lower thresholds. For example, the highest default quality filter of SpectraST employs a threshold of two replicate spectra (20). Additionally, one would expect that the number of MCIPs is dependent on the number of replicates, if the phenomenon would be caused by random fluctuations. Indeed, the number of replicate spectra is virtually uncorrelated with the number of CIPs (Fig. 2c). Thus, we can strengthen our point that MCIPs are a stable phenomenon that is not caused by random divergences or impurities.

Variation among fragmentation spectra depending on machine and fragmentation type has recently been examined (43) and it is well known that peptide fragmentation is subject to intrinsic variation. As the total number of detected ions underlying one fragmentation spectrum is large, it is reasonable to assume, that this variation is not stochastic in nature, but rather determined by experimental conditions. The method proposed in our study tries to capture these underlying deterministic variations. We speculate that the MCIPs we observe are boundary cases that characterize the range of possible fragmentation spectra. Hence, our approach of MCIPs determined by a systematic clustering method might be a solution to the common problem that spectral libraries are highly dependent on experimental conditions (44). This is underlined by the fact, that we are able to obtain almost complete coverage (Fig 3a) of an extremely heterogeneous mixture of peptides stemming from three different species, measured in over a hundred of different LC-MS/MS runs over many months. The observation that the number of false positives is only marginally affected upon integration of MCIPs (Fig. 4c) validates that this coverage can be achieved at virtually the same error rate.

For practical purposes, the fact that the average number of different clusters is about two and no new clusters appear after assembling about 20 replicate patterns (Fig. 2d) indicates, that the phenomenon is easily manageable for the creation of a library.

Our reductionist approach of using MaxQuant preprocessed spectra comes at the cost of possibly neglecting important spectral information. The dot score values determined from this approach will be different to the dot score values derived from raw spectra, as the representative vectors are shorter and vector length influences the outcome. Nevertheless, heuristic measures to shorten the vector are applied in common library generation tools (20, 22) and have shown to only mildly affect the overall sensitivity. Additionally, we validated the similarity threshold used in this study to be discriminative on our spectra (supplemental Fig. S3). Additionally, we carried out analyses on raw, or semi raw spectra at different ppm precisions using MaxQuant processed .apl data and MS-Convert processed .mzXML profile data and see systematically lower dot scores as well as abundant MCIPs under all conditions (supplemental Fig. S6).

The large increase in spectral recognition we see in our approach compared to SpectraST is likely due to the employed test set, which contains at least 10 replicate spectra for each peptide and is not the usual application scenario for a spectral search engine. Nevertheless, it is a valid performance test and the results imply, that choosing the best CIP is crucial.

With the advent of quantitative DIA methods like SWATH, the phenomenon of MCIPs becomes important in the context of quantification. If MCIPs are not taken into account, in a significant fraction of cases, fold changes might be drastically miscalculated because peptides that are actually there, will be missed because the fragmentation spectrum is different. The increase in spectral recognition of our SWATH data set upon integration of MCIPs (Fig. 6) is a first indication that SWATH benefits from our approach. Additionally, it has recently been shown that SWATH data acquisition is improved when local libraries are used in addition to public libraries (44). As this might very well be, because the intrinsic variation of the machine is better covered this way, employing the MCIP approach in public spectral libraries could eliminate the need for context-dependent libraries.

The MCIP approach contributes a simple and pragmatic solution to the problem of variable peptide fragmentation spectra. We emphasize that even in the current time of ever more precise machines and measurement possibilities, peptide fragmentation can still be subject to strong, but reproducible, fluctuations. Ideally, this problem may be solved by new generations of machines producing unique fragmentation patterns. In the meantime, the use of MCIPs in spectral libraries is a simple and sufficient solution.

References

1. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature*, *422*(6928), 198-207

2. Domon, B., and Aebersold, R. (2006) Mass spectrometry and protein analysis. *Science*, *312*(5771), 212-217

3. Steen, H., and Mann, M. (2004). The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.*, *5*(9), 699-711

4. Michalski, A., Cox, J. and Mann, M. (2011) More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC−MS/MS. *J. Proteome Res.*, 10(4), 1785-1793

5. Liu, H., Sadygov, R. G., and Yates, J. R., 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal. Chem. 76, 4193–4201

6. Schmidt, A., Gehlenborg, N., Bodenmiller, B., Mueller, L. N., Campbell, D., Mueller, M., Aebersold, R., and Domon, B. (2008) An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. *Mol. Cell. Proteomics* 7, 2138–2150

7. Picotti, P., Aebersold, R. (2012) Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* 6, 555-566

8. Purvine, S., Eppel, J. T., Yi, E. C., and Goodlett, D. R. (2003) Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* 3, 847–850

9. Plumb, R. S., Johnson, K. A., Rainville, P., Smith, B. W., Wilson, I. D., Castro-Perez, J. M., and Nicholson, J. K. (2006) UPLC/MS(E): A new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Commun. Mass Spectrom.* 20, 1989–1994

10. Geiger, T., Cox, J., and Mann, M. (2010) Proteomics on an Orbitrap bench- top mass spectrometer using all ion fragmentation. *Mol. Cell. Proteomics* 9, 2252–2261

11. Venable, J. D., Dong, M. Q., Wohlschlegel, J., Dillin, A., and Yates, J. R. (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* 1, 39–45

12. Panchaud, A., Scherl, A., Shaffer, S. A., von Haller, P. D., Kulasekara, H. D., Miller, Miller, S. I., and Goodlett, D. R. (2009) Precursor acquisition independent from ion count: How to dive deeper into the proteomics ocean. *Anal. Chem.* 81, 6481–6488

13. Bern, M., Finney, G., Hoopmann, M. R., Merrihew, G., Toth, M. J., and MacCoss, M. J. (2010) Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Anal. Chem.* 82, 833–841

14. Carvalho, P. C., Han, X., Xu, T., Cociorva, D., Carvalho Mda, G., Barbosa, V. C., and Yates, J. R., 3rd (2010) XDIA: Improving on the label-free data-independent analysis. *Bioinformatics* 26, 847–848

15. Panchaud, A., Jung, S., Shaffer, S. A., Aitchison, J. D., and Goodlett, D. R. (2011) Faster, quantitative, and accurate precursor acquisition independent from ion count. *Anal. Chem.* 83, 2250–2257

16. Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R. and Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol.*

Cell. Proteomics, *11*(6), O111-016717

17. Röst, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S. M., Schubert, O. T., Wolski W., Collins B.C., Malmström J., Malmström L. and Aebersold, R. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.*, *32*(3), 219-223

18. Tsou, C. C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A. C., and Nesvizhskii, A. I. (2015). DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods*, *12*(3), 258-264

19. Li Y, Zhong CQ, Xu X, Cai S, Wu X, Zhang Y, Chen J, Shi J, Lin S, Han J. (2015). Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. *Nat. Methods*, *12*(12), 1105-1106

20. Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., Stein, S. E., and Aebersold, R. (2008). Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods*, *5*(10), 873-875

21. Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S., and MacCoss, M. J. (2006). Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.*, *78*(16), 5678-5684

22. Craig, R., Cortens, J. C., Fenyo, D., and Beavis, R. C. (2006). Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.*, *5*(8), 1843-1849

23. Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., King, N., Stein, S. E., and Aebersold, R. (2007). Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*, *7*(5), 655-667

24. Stein, S. E., and Scott, D. R. (1994) Optimization and Testing of Mass- Spectral Library Search Algorithms for Compound Identification. *J. Am. Soc. Mass Spectrom.* 5, 859–866

25. Beer, I., Barnea, E., Ziv, T., and Admon, A. (2004) Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* 4, 950–960

26. Sherwood, C. A., Eastham, A., Lee, L. W., Risler, J., Vitek, O., and Martin, D. B. (2009) Correlation between y-Type Ions Observed in Ion Trap and Triple Quadrupole Mass Spectrometers. *J. Proteome Res.* 8, 4243–4251

27. Yen, C. Y., Houel, S., Ahn, N. G., and Old, W. M. (2011) Spectrum-to- Spectrum Searching Using a Proteome-wide Spectral Library. *Mol. Cell. Proteomics* 10, M111.007666

28. Wan, K. X., Vidavsky, I., and Gross, M. L. (2002) Comparing similar spectra: from similarity index to spectral contrast angle. *J. Am. Soc. Mass Spectrom.* 13, 85–88

29. Tabb, D. L., MacCoss, M. J., Wu, C. C., Anderson, S. D., and Yates, J. R. (2003) Similarity among tandem mass spectra from proteomic experiments: Detection, significance, and utility. *Anal. Chem*. 75, 2470–2477

30. Wang, J. A., Perez-Santiago, J., Katz, J. E., Mallick, P., and Bandeira, N. (2010) Peptide Identification from Mixture Tandem Mass Spectra. *Mol. Cell. Proteomics* 9, 1476–1485

31. Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., and Aebersold, R. (2006) The PeptideAtlas project. *Nucleic Acids Res.* 34, D655–D658

32. Jones, P., Cote, R. G., Martens, L., Quinn, A. F., Taylor, C. F., Derache, W., Hermjakob, H., and Apweiler, R. (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.* 34, D659–D663

33. Riffle, M., and Eng, J. K. (2009) Proteomics data repositories. Proteomics 9, 4653–4663

34. Vizcaino, J. A., Cote, R. G., Csordas, A., Dianes, J. A., Fabregat, A., Foster, J. M., Griss, J., Alpi, E., Birim, M., Contell, J., O'Kelly, G., Schoenegger, A., Ovelleiro, D., Perez-Riverol, Y., Reisinger, F., Rios, D., Wang, R., and Hermjakob, H. (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* 41, D1063–D1069

35. Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, *26*(12), 1367-1372

36. Cox, J., Neuhauser N., Michalski A., Scheltema R. A., Olsen, J.V., and Mann M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. proteome res.*, 10(4), 1794-1805

37. Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun Z., Nilsson E., Pratt B., Prazen B., Eng, J. K., Martin D.B., Nesvizhskii A.I. and Aebersold R. (2010). A guided tour of the Trans-Proteomic Pipeline. *Proteomics*, *10*(6), 1150-1159

38. Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M. Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S. L., Nuwaysir, L. M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E. W., Moritz, R. L., Katz, J. E., Agus, D. B., MacCoss, M., Tabb, D. L., and Mallick, P. (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* 30, 918–920

39. Craig, R., and Beavis, R. C. (2004). TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, *20*(9), 1466-1467

40. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, *74*(20), 5383-5392

41. Lam, H., Deutsch, E. W., Aebersold, R. (2009). Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. *J. proteome res.*, 9(1), 605-610

42. Griss, J. (2016). Spectral library searching in proteomics. *Proteomics*, 16: 729–740

43. Toprak, U. H., Gillet, L. C., Maiolica, A., Navarro, P., Leitner, A., and Aebersold, R. (2014). Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a

discriminating feature for targeted proteomics. *Mol. Cell. Proteomics*, *13*(8), 2056-2071

44. Wu, J. X., Song, X., Pascovici, D., Zaw, T., Care, N., Krisp, C. and Molloy, M. P. (2016). SWATH mass spectrometry performance using extended peptide MS/MS assay libraries. *Mol. Cell. Proteomics*, 15, 2501-2514

45. Gallien, S., Duriez, E., Demeure, K., & Domon, B. (2013). Selectivity of LC-MS/MS analysis: implication for proteomics experiments. *J. prote*omics, 81, 148-158.

Footnotes

C.A. is supported by a DFG fellowship through QBM. We acknowledge partial funding by the SFB 1123 of the DFG and by the BioSysNet research network. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Figure Legends

Table 1: Overview of different library generation workflows and comparison of main features. Listed are Andromeda (36), a sequence based search engine, the multiple characteristic intensity patterns (MCIP) approach of this paper and the SpectraST (23) approach.

Fig. 1: Schematic display of the decoy spectra approach. A cross validation test set is generated, containing 1:1 decoy and original spectra (left). The original spectra not present in the test set are then used to extract characteristic intensity patterns (CIPs), which are used to search the test set (right).

Fig. 2: Analysis of multiple characteristic intensity patterns (MCIPs). a) The percentage of peptide charge pairs (10.580 total) containing confidently identified MCIPs for different rematch constraints. The legend indicates the minimum number of spectra matching with the second CIP. b) Normalized ion intensities (NRFVs) of the two largest spectral clusters (neighborhoods) plotted on top of each other for two selected peptide charge pairs. Blue coloring indicates spectra belonging to the largest cluster 1, red coloring indicates spectra belonging to cluster 2. This demonstrates, that fragmentation spectra can differ for the same peptide charge pair (clearly different fragmentation characteristics between blue and red) but are individually reproducible (similarity of fragmentation characteristics within blue and red). The line plots are used to better show individual differences between the spectra and an annotated format of the spectra displayed is visible in supplemental figure S11. c) Dependence of the number of clusters on the number of replicate spectra (left: all peptides, right: only MCIP peptide charge pairs). Even for hundreds of replicate spectra, the number of CIPs rarely exceeds six, indicating saturation for the number of CIPs.

Fig. 3: a) Spectral coverage depending on the number of characteristic intensity patterns (CIPs) integrated in the spectral library. The number of peptide charge pairs (y-axis) is displayed, for which the spectral coverage is larger or equal to the value denoted on the x-axis in percent. The single CIP approach (red line) leaves a large fraction of spectra uncovered. Integrating one more CIP (blue line) into the library gives a drastic increase in coverage with successively smaller increases upon integration of more CIPs until almost complete coverage is reached. b) Cumulative distribution of the relative fractions of the clusters among all replicate spectra.

Fig. 4: Spectral coverage of the SpectraST search engine (dark blue line) in comparison with the MCIP approach (red to green). A single characteristic intensity pattern (CIP) still yields partially lower coverage (red line), while upon integration of the second CIP almost complete coverage is achieved.

Fig. 5: Assessment of accuracy using randomized decoy spectra, as described in the methods section, cumulative depiction. a) Comparison of the overall identification accuracy between the single characteristic intensity pattern (CIP) approach (red line) and multiple CIPs (MCIPs) (blue line, green line). A significant improvement upon integration of MCIPs is visible. b) Effect of MCIP integration on false negatives: The false negatives rate is strongly decreased (green and blue line) as now also the differently fragmenting ions are integrated. c) Effect of MCIP integration on the false positives rate: The plot demonstrates that the increase in recognition does not incur significant loss of specificity, as the false positive rate is only marginally increased upon integration of MCIPs (green and blue line).

Fig. 6: Application of the MCIP approach on SWATH data and comparison with DDA data. For reasonable similarity thresholds, a significant decrease in unidentified peptides can be seen on

SWATH data when integrating MCIPs (violet line) in comparison to a single CIP (green line). An analogous behavior is seen for the DDA approach, with a significantly smaller number of missed peptides (blue line MCIPs, red line single CIP).

Tables

Table 1

Andromeda (MaxQuant)	MCIP	SpectraST
Library Creation		
• library contains all peptide sequences in a sequence database • calculate theoretical fragment masses (b- and y- ions with modifications, modification specific neutral losses, diagnostic peaks)	Data Selection	
	• library built from previously measured charged peptides • obtain all measured spectra of a peptide charge pair, same modification (= replicate spectra)	
	• import MaxQuant identified spectra • MaxQuant preprocessed • 1% FDR • only b- and y- ions (w. losses) • no ranking, or ranked by signal to noise	• import PeptideProphet ranked spectra • not preprocessed • PeptideProphet score >0.9 • all signals in spectrum • ranked by signal to noise
	Clustering	
	• assess similarities between all replicate spectra • iteratively select either largest clusters or representative clusters of similar spectra	• assess similarities for rank-selected spectra • select representative (not overall largest) cluster
	Cluster Processing	
	• select centroid spectrum of each cluster • create a minimum set of multiple characteristic patterns {CIP_1, CIP_2 ,...CIP_n}	• keep only the representative cluster • align peaks • select consensus peaks (also non-canonical) • normalize and average representative cluster • create a single SpectraST characteristic pattern {CIP}
Searching		
• MaxQuant preprocessed query spectra (denoising, mass-centroiding, de-isotoping, etc..)		• unprocessed query spectra (e.g. .mzXML)
• match just masses, no intensities	• match masses and intensities	

Figures

Figure 1

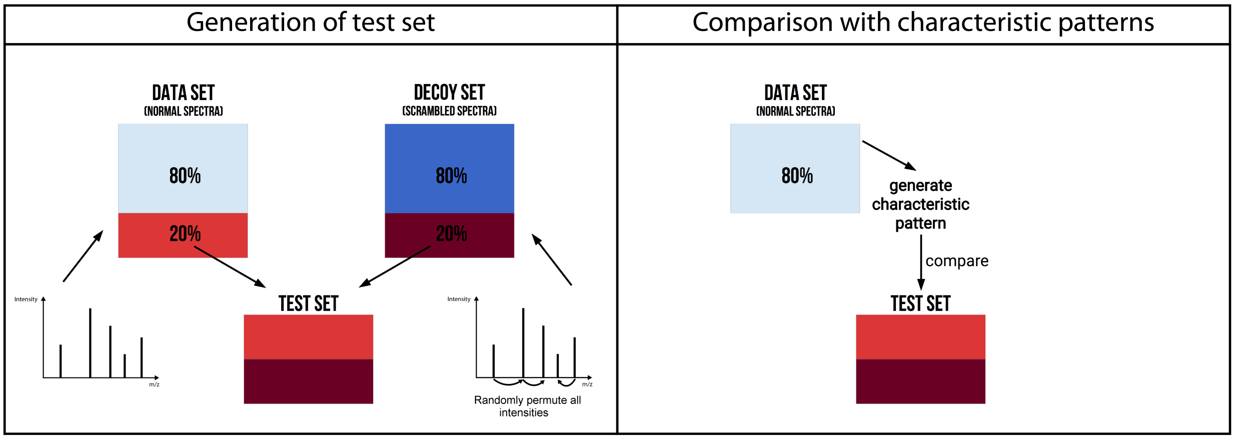
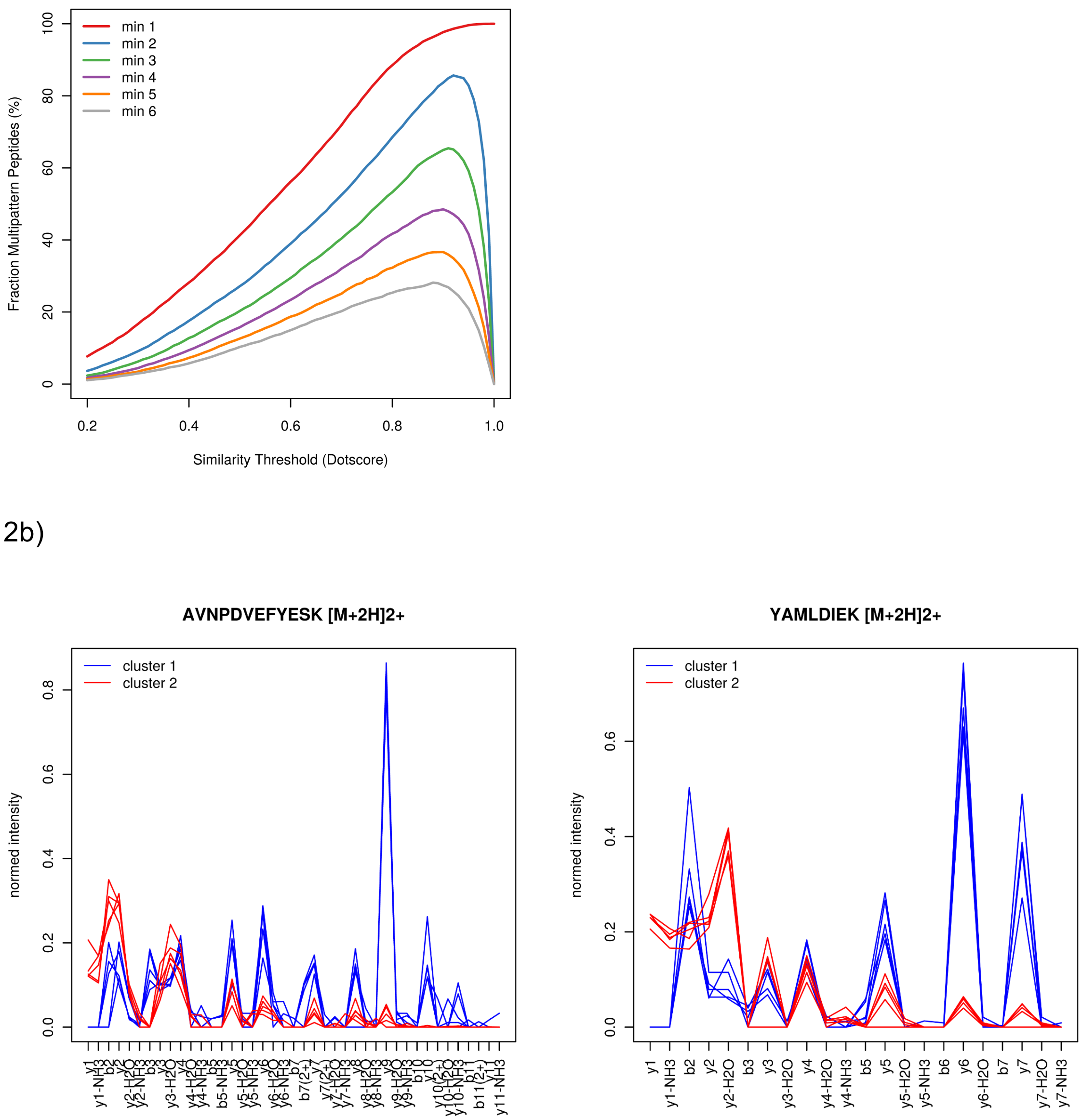


Figure 2



2a)

2b)

2c)

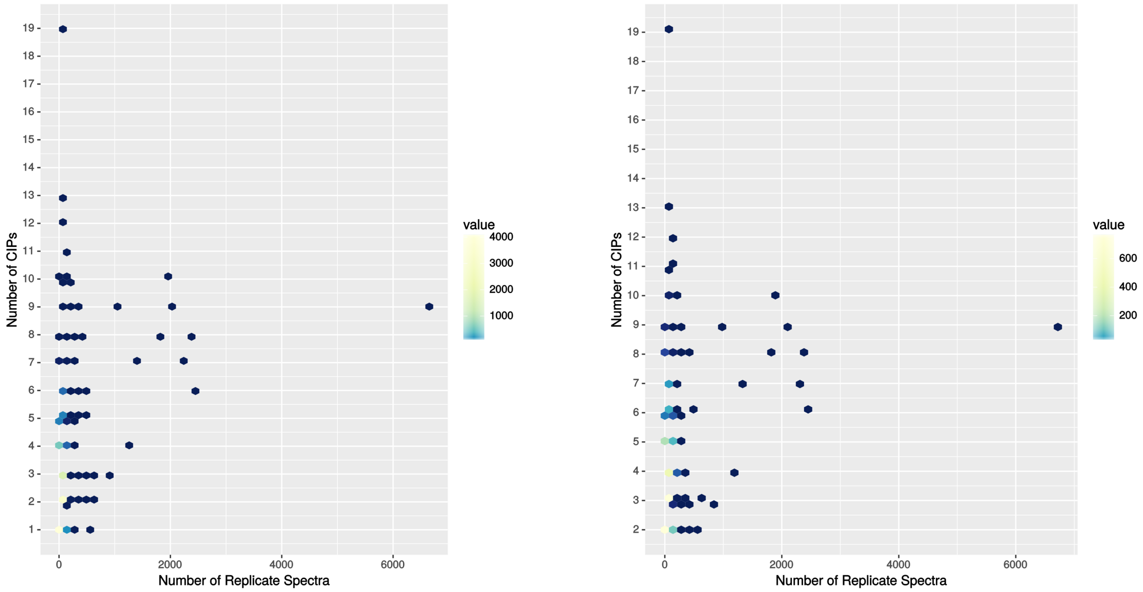
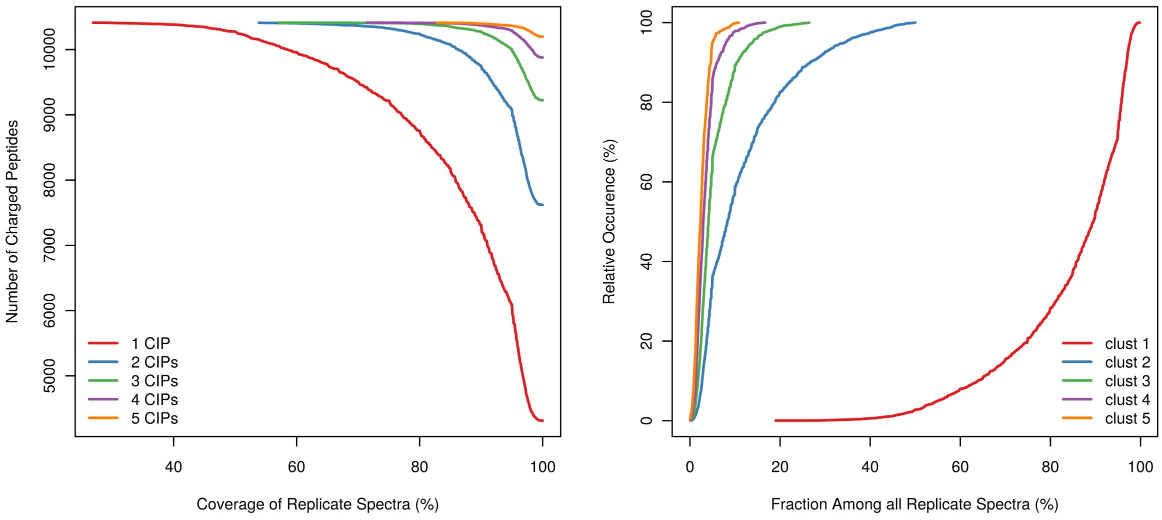

Number of CIPs
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
2000
4000
6000
Number of Replicate Spectra
value
4000
3000
2000
1000
Number of CIPs
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
0
2000
4000
6000
Number of Replicate Spectra
value
600
400
200

Figure 3


3a)

3b)

Figure 4

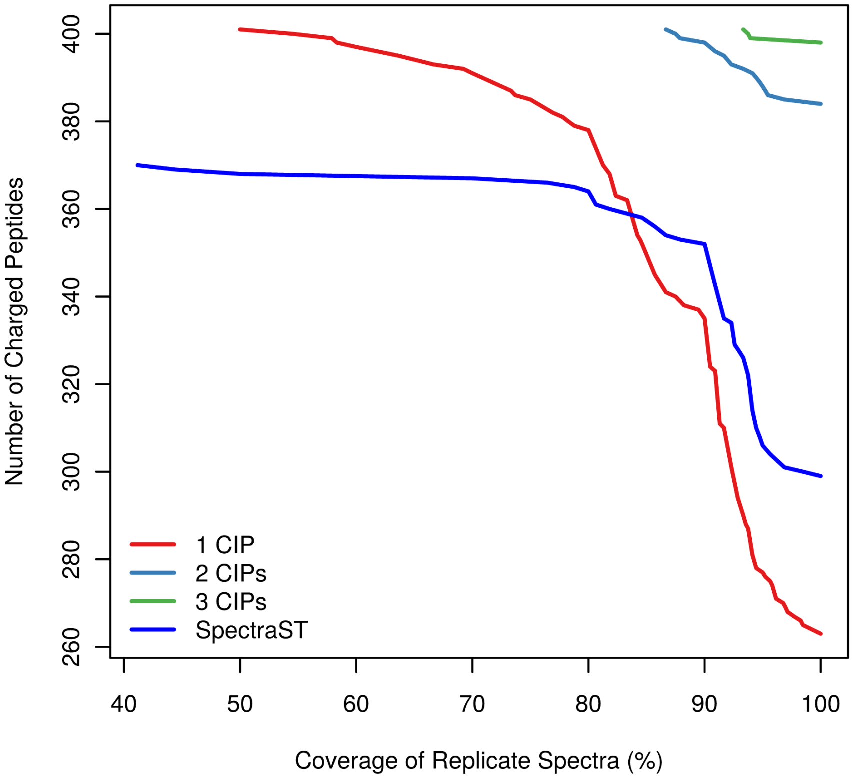

Number of Charged Peptides
260
280
300
320
340
360
380
400
40
50
60
70
80
90
100
Coverage of Replicate Spectra (%)
1 CIP
2 CIPs
3 CIPs
SpectraST

Figure 5

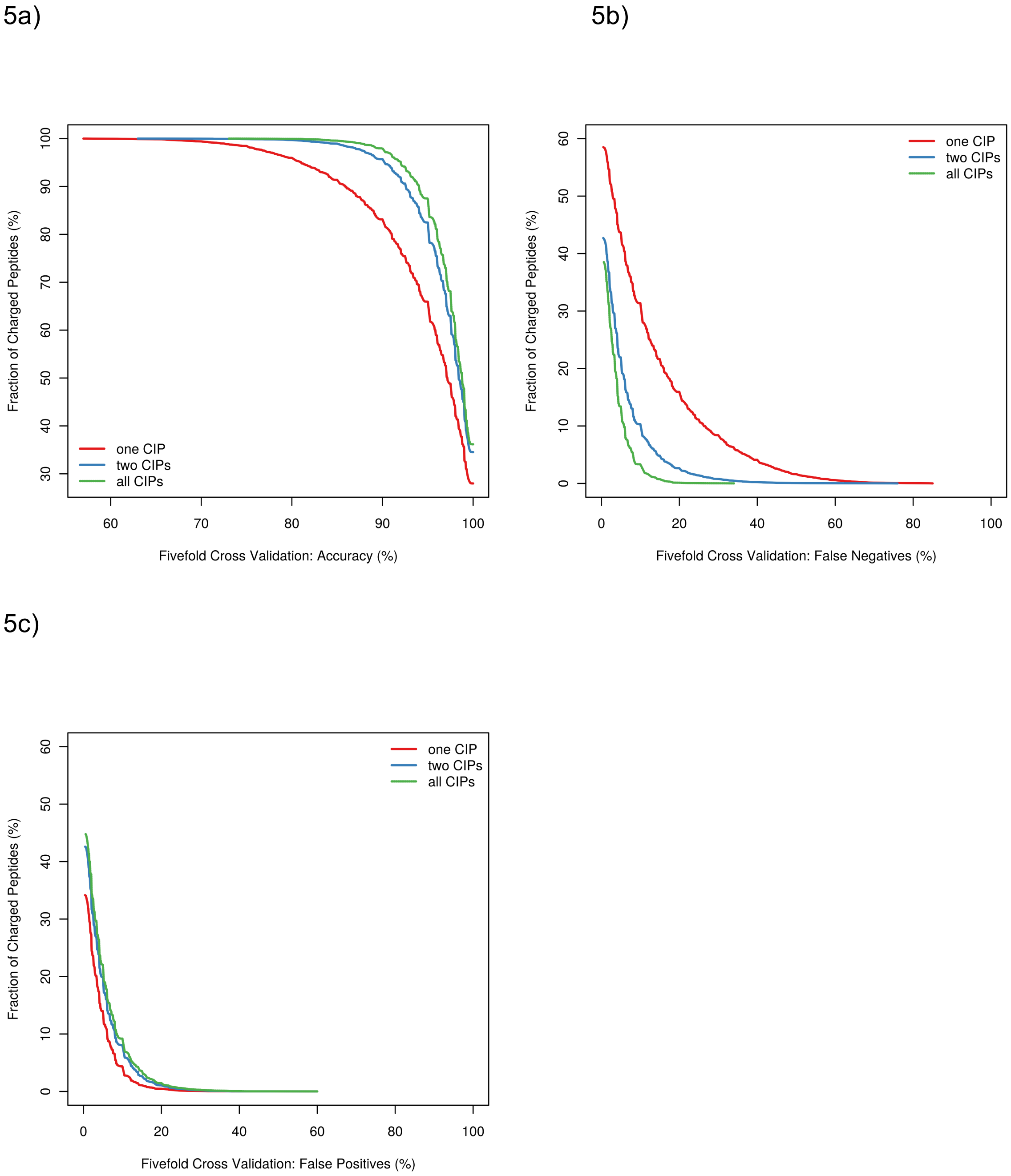
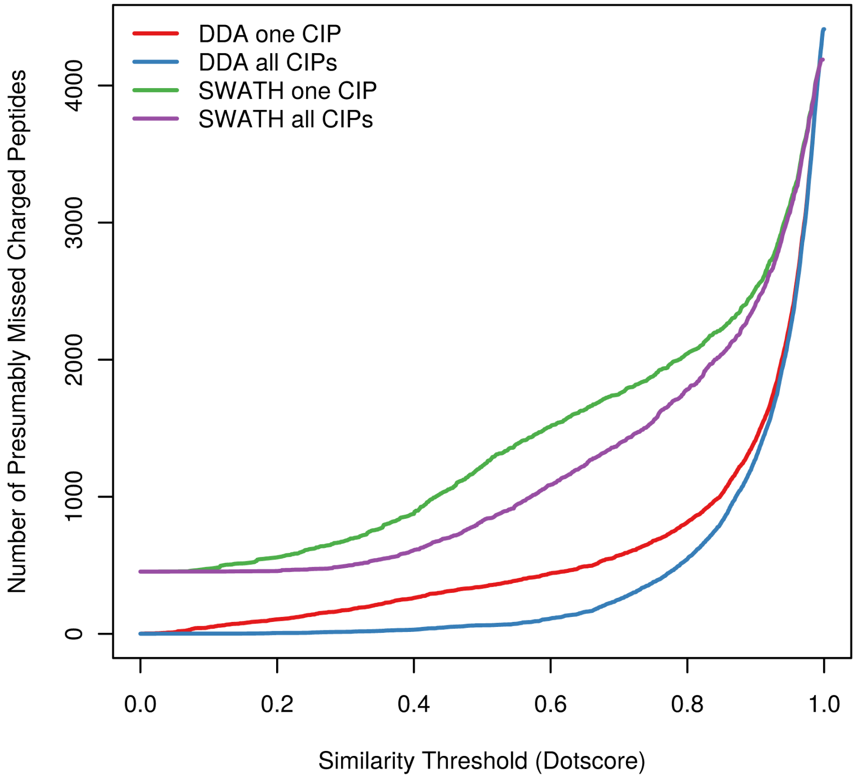

5a)
Fraction of Charged Peptides (%)
100
90
80
70
60
50
40
30
60
70
80
90
100
Fivefold Cross Validation: Accuracy (%)
one CIP
two CIPs
all CIPs
5b)
Fraction of Charged Peptides (%)
60
50
40
30
20
10
0
0
20
40
60
80
100
Fivefold Cross Validation: False Negatives (%)
one CIP
two CIPs
all CIPs
5c)
Fraction of Charged Peptides (%)
60
50
40
30
20
10
0
0
20
40
60
80
100
Fivefold Cross Validation: False Positives (%)
one CIP
two CIPs
all CIPs

Figure 6


DDA one CIP
DDA all CIPs
SWATH one CIP
SWATH all CIPs
Number of Presumably Missed Charged Peptides
0
1000
2000
3000
4000
0.0
0.2
0.4
0.6
0.8
1.0
Similarity Threshold (Dotscore)

Equations

Equation 1:

$$I = (i_1, \ldots, i_n).$$

Equation 2:

$$DP(X,Y) = \sum_{k=1}^{n} x_k \cdot y_k.$$