Building high-quality assay libraries for targeted analysis of SWATH MS data

Schubert, O. T., Gillet, L. C., Collins, B. C., Navarro, P., Rosenberger, G., Wolski, W. E., Lam, H., Amodei, D., Mallick, P., MacLean, B., & Aebersold, R. (2015). Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nature Protocols*, *10*(3), 426-441. https://doi.org/10.1038/nprot.2015.015

Published in:
Nature Protocols

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Building high-quality assay libraries for targeted analysis of SWATH MS data

Olga T. Schubert[1,2,*], Ludovic C. Gillet[1,*], Ben C. Collins[1,*], Pedro Navarro[3], George Rosenberger[1,2], Witold E. Wolski[1,4], Henry Lam[5], Dario Amodei[6], Parag Mallick[6], Brendan MacLean[7], and Ruedi Aebersold[1,8,**]

1 Institute of Molecular Systems Biology, Department of Biology, ETH Zurich, Zurich, Switzerland
2 PhD Program in Systems Biology, University of Zurich and ETH Zurich, Zurich, Switzerland
3 Institute for Immunology, University Medical Center of the Johannes-Gutenberg University Mainz, Mainz, Germany
4 SystemsX.ch Biology IT (SyBIT), SystemsX.ch, Zurich, Switzerland
5 Division of Biomedical Engineering and Department of Chemical and Biomolecular Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China
6 Department of Radiology, Stanford University School of Medicine, Stanford, California, USA
7 Department of Genome Sciences, University of Washington, Seattle, Washington, USA
8 Faculty of Science, University of Zurich, Zurich, Switzerland

* These authors contributed equally to this work.
** Correspondence to Ruedi Aebersold (aebersold@imsb.biol.ethz.ch)

Abstract

Targeted proteomics by selected/multiple reaction monitoring (S/MRM) or, on a larger scale, by SWATH MS typically relies on spectral reference libraries for peptide identification. Quality and coverage of these libraries are therefore of critical importance for the performance of the methods. Here we present a detailed protocol that has been successfully used to build high-quality, extensive reference libraries supporting targeted proteomics by SWATH MS. We describe each step of the process, including data acquisition by discovery proteomics, assertion of peptide-spectrum matches, generation of consensus spectra and compilation of mass spectrometric coordinates that uniquely define each targeted peptide. Crucial steps such as FDR control, retention time normalization and handling of post-translationally modified peptides are detailed. Finally we show how to use the library to extract SWATH data with the open-source software Skyline. The protocol takes 2-3 days to complete, depending on the extent of the library and the computational resources available.

INTRODUCTION

Most proteomic analyses involve one or several of an array of mass spectrometric (MS) methods. To date, the most frequently used method is data-dependent acquisition (DDA) because of its unmatched capabilities to identify the protein components of a sample (see **Box 1** for brief explanations on relevant proteomics terminology). DDA-based methods, also referred to as shotgun proteomics, have been widely used to identify and quantify proteins. However, in highly complex proteomic samples the semi-stochastic nature of DDA leads to some curtailments in the consistency of quantification across many samples, particularly for lower abundant peptide species[1]. More recently, mainly driven by the demands of translational research and systems biology projects, the need to generate data which allow the comparative relative or absolute quantification of sets of proteins, reproducibly and accurately across sample cohorts numbering tens to hundreds of samples has been recognized. At present, the method of choice for such studies is targeted mass spectrometry (MS) where definitive assays are used to quantify pre-determined sets of proteins across samples at a high degree of reproducibility. The gold standard targeting MS technique is selected reaction monitoring (SRM), also referred to as multiple reaction monitoring (MRM)[2]. In 2012, we introduced SWATH MS[3] as a next generation targeting method that largely maintains the favourable performance characteristics of S/MRM such as quantitative accuracy, dynamic range, and reproducibility, while substantially extending the number of quantifiable peptides from the range of tens to hundred (with scheduling) per sample injection for S/MRM to thousands or tens of thousands per sample injection for SWATH MS. Thus, SWATH MS supports the accurate relative quantification of large fractions of a proteome in a single injection[3].

SWATH MS is a variant of the class of data-independent acquisition (DIA) methods that record fragment ion spectra of all ionized species of a sample[4,5]. For SWATH MS data acquisition, a high-resolution quadrupole-time-of-flight mass spectrometer cycles through a series of fixed precursor isolation windows that collectively cover the entire m/z range of MS-suitable peptides and acquires composite fragment ion spectra from all the precursor ions contained in a specific window at a given time. The window size and dwell time are chosen such that the cycle time is short enough to allow each peptide to be fragmented approximately 8-10 times across its chromatographic elution profile. A SWATH MS dataset therefore constitutes a complete digital record of all ionized species above the detection limit where the fragment ion spectra of individual peptides are represented in a convoluted, but highly structured manner. The quality of these digital maps mainly depends on the precursor isolation window width, fragment ion resolution, dwell and cycle time. To identify and quantify peptides in such SWATH MS fragment ion maps we have devised a targeted data analysis strategy[3] that is supported by software tools such as OpenSWATH[6], PeakView (AB Sciex), Spectronaut[7], or Skyline[8] and is modelled after the automated identification of peptides by

S/MRM[9]. In essence, these tools identify peak groups that uniquely associate with the targeted peptide within the comprehensive SWATH MS signal map, and then compute a probability that the targeted peptide has been correctly identified. The peak groups consist of the signals of specific fragment ions derived from the target peptide (transitions) integrated over chromatographic time. The set of transition signals that identifies a target peptide with the highest sensitivity and specificity constitutes a definitive assay for the detection of that peptide and has to be determined prior to the analysis.

A high-quality library of assays is a prerequisite for SWATH MS and similar targeting MS methods[10]. Such an assay library is typically built from compendia of fragment ion spectra (spectral libraries) and contains the exact mass spectrometric coordinates for each targeted peptide. For each peptide, these coordinates consist of (i) the peptide precursor m/z, (ii) the m/z for a selection of its fragment ions together with their relative intensities, and (iii) the chromatographic retention time of the peptide in a normalized retention time space. Ideally, the peptides in the assay library cover all proteins of interest for a particular study, or even an entire proteome.

Over the past years we and others have developed software tools for the generation of spectral libraries (SpectraST[11], X!Hunter[12], Bibliospec[13]). They were originally devised for searching DDA datasets by spectral matching. Analogous spectral libraries have also been used for targeted proteomics by S/MRM[8,14-16] or SWATH MS[3,17], ideally built from fragment ion spectra generated on the same type of instrument used for targeting. To eliminate the need for assay generation for each experiment, our group has spearheaded the development of publicly accessible assay libraries for the entire proteome of *Saccharomyces cerevisiae*[18], *Mycobacterium tuberculosis*[19] and *Streptococcus pyogenes*[20], as well as for disease relevant human subproteomes, including the human glycoproteome[21] and a set of cancer associated proteins[22]. Most of these assay libraries were optimized for S/MRM and are available through the SRMAtlas database (www.SRMAtlas.org). More recently we developed an assay library optimized for SWATH MS which covers more than 10,000 human proteins annotated in the UniProtKB/SwissProt database[23].

In this paper we describe a step-by-step protocol and an integrated, openly accessible computational pipeline to generate high-quality assay libraries for targeted MS. All required software tools are freely available through the TPP[24], ProteoWizard[25], and OpenMS[26] software suites or provided as a python package together with this protocol. For the purpose of user friendliness we implemented the protocol on a Windows platform. The computational pipeline described here allows maximal control over each step of the library building process and is suitable for large, organism-wide assay libraries as well as for experiment-specific assay libraries generated from as few as a single DDA dataset. The conceptual workflow and considerations to be made at each step are, however, generic and other tools such as the integrated ProteinPilot-PeakView pipeline, Spectronaut (supporting MaxQuant[27] search

engine outputs), or Skyline (supporting various search engine outputs) might be used instead. These integrated pipelines allow less control over the workflow, but nevertheless might provide a suitable alternative for researchers who wish to avoid data handling by command line tools and python scripts.

The assay library building workflow described here is optimized for SWATH MS. However, in combination with dedicated analysis tools, it is also applicable to other targeted MS techniques, including S/MRM. Moreover, many of the considerations described here are also valid for building libraries in the context of spectral library searching of DDA[28,29] or DIA[30] data sets.

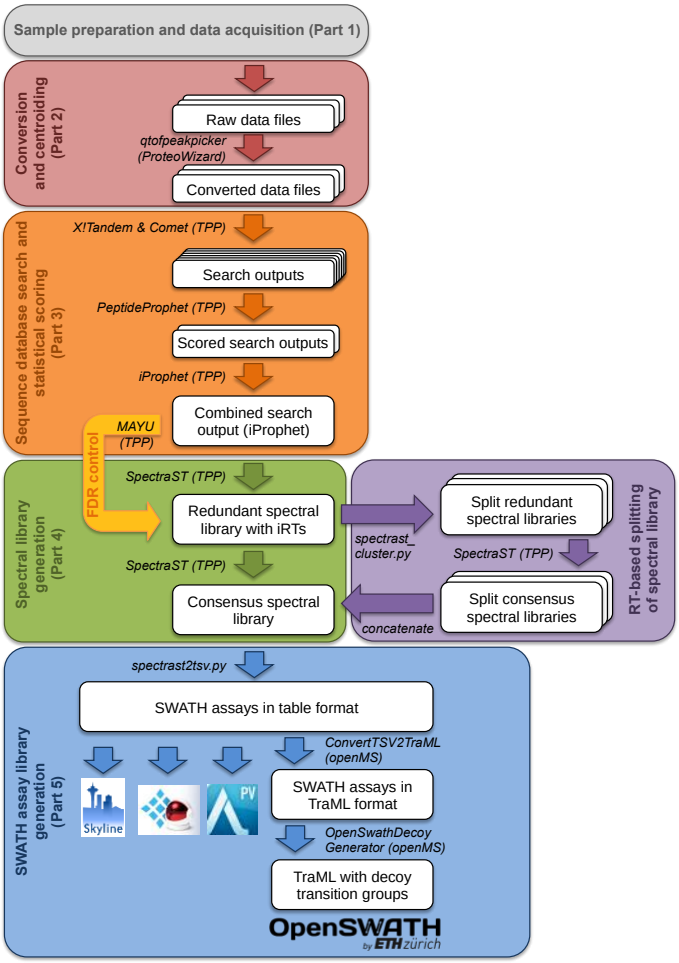

Figure 1 | Workflow for SWATH assay library generation. The library building workflow starts with the selection of representative samples and fragment ion spectra acquisition (Part 1), followed by centroiding and conversion of the raw files into an open format (Part 2). The centroided fragment ion spectra are searched against a protein sequence database to establish PSMs (Part 3). The confidently assigned spectra are then converted into a spectral library and all retention times are normalized and converted into iRTs (Part 4). To address complications when building libraries for post translationally modified peptides, an optional subroutine has been developed to account for potential errors in site localization of modifications. After consensus library generation the most intense fragment ions of each peptide precursor are selected (Part 5). Optionally the resulting assay library in table format can be converted into TraML format and decoy transition groups can be added if required for downstream analysis.

The protocol in overview.

The purpose of this protocol is to build assay libraries supporting targeted analysis of SWATH MS proteomic datasets. The protocol covers the acquisition of high quality fragment ion spectra in DDA mode, the assignment of peptide sequences to these spectra, their conversion into spectral libraries, and the compilation of the final assays from the spectral

libraries. The steps required for custom SWATH assay library generation are outlined in Figure 1 and described in detail in the following paragraphs. In the section “Procedure” we provide detailed step-by-step instructions, including a detailed list of all files produced during the workflow (Supplementary Note 1), installation and usage of all software tools (Supplementary Note 2 and 3), on a prototypical dataset consisting of three samples from a yeast osmotic shock time course experiment (unpublished data). In **Box 2** we provide a checklist summarizing the most critical points of the protocol.

Part 1a (Step 1 of the Procedure): Selection of representative samples for the generation of spectral libraries

Peptides can only be identified by targeted MS if they are included in the assay library. To cover proteins that are expressed in specific biological conditions the samples used to generate those assay libraries should therefore represent the entire biological space to be quantified by SWATH MS. Also, because SWATH MS analysis is more sensitive in identifying and quantifying peptides compared to DDA in side-by-side analyses on the same instrument[3], it might be worth fractionating the samples used to generate the assay library prior to the analysis e.g. by using isoelectric focusing by off-gel electrophoresis[31], or SDS page[10]. The resulting fractions are then subjected to DDA, preferably on the same type of instrument that is also used for the subsequent SWATH MS analyses.

As an alternative to the use of fragment ion spectra of native peptides, assay libraries can be built from or supplemented with synthetic peptides[16] or recombinant proteins[32], or computationally predicted. We have previously shown that unpurified unlabelled synthetic peptides produce spectra which are indistinguishable from those derived from natural endogenous peptides[18]. By analysing pools of 100 – 1000 synthetic peptides, high-quality fragment ion spectra can be produced very efficiently for large numbers of target proteins, even if they have never been observed from a natural source[18,19]. A number of approaches and tools have been described to predict those peptides of a protein that are most suitable for targeted analysis (PeptideRank[33], PeptideSieve[34], CONSeQuence[35], ESPPredictor[36], Detectability Predictor[37], STEPP[38]), though their predictors are less accurate than empirical assessment of optimal peptides[32]. Fragment ion spectra can also be entirely computationally predicted using physicochemical models of peptide fragmentation or by machine learning approaches. Predicted fragment ion spectra, however, are also expected to less faithfully match empirical spectra of native peptides[39]. For library building, empirical data, either from native or synthetic peptides, are therefore preferable over computationally predicted spectra.

Generally, it has been shown that very extensive SWATH assay libraries, for example those resulting from fractionation prior to MS analysis and/or addition of synthetic peptides, lead to more peptide and protein identifications during SWATH data analysis without impairing quantification accuracy[23]. It is, however, important to note that such extensive assay libraries require more stringent FDR control during the SWATH data analysis because the increased search space results in higher numbers of false positive identifications (see also **Box 3**)[23].

Incidentally, the presence of an assay in the library does not guarantee that the corresponding peptide can be detected i.e. if the peptide is not present or below the limit of detection in the SWATH analysis of a given sample.

Part 1b (Step 2): Working with retention time reference peptides

Because the chromatographic retention time of the targeted peptides is an essential component of the final peptide assay it is recommended to work with retention time reference peptides[40] which are spiked into all samples that are used for library generation. This will allow effective peptide retention time normalization and retention time-based splitting of fragment ion spectra, to determine most accurate retention times for each targeted peptide. Alternatively, endogenous retention time reference peptides can be used for retention time normalization[10]. As with spike-in reference peptides, endogenous reference peptides need to cover a large retention time range of the sample and need to be well detectable over all samples. All downstream steps are identical and the quality of the alignment as well as the resulting data is very comparable between the workflow with spike-in and endogenous retention time reference peptides. As a reference scale for the retention time normalization either the unit-less iRT scale[40] or the retention times in minutes from any previously acquired MS injection can be chosen. The main advantage of using the iRT scale is that it is a defined reference and therefore facilitates transferability between instruments and labs.

Part 1c (Step 3): Acquisition of fragment ion spectra

Optimal performance of the assay library for the targeted identification of peptides in SWATH MS datasets is achieved if the spectra used to generate the spectral libraries reflect as closely as possible the relative fragment ion intensities in the SWATH MS maps. To ensure optimal portability of the assays it is therefore highly recommended to use the same type of instrument for library generation as for SWATH MS analysis[41]. If no TripleTOF 5600 mass spectrometer is available to generate the library, an alternative instrument with beam-type collision cell or ion trap-type collision cell operable in HCD mode may be used instead, as they generate fragmentation patterns that are similar to those generated by the TripleTOF 5600 instrument[42]. If data sources other than those from a TripleTOF 5600 instrument are used, it is recommended to ensure that the relative fragment ion intensities do not exceed 30% variation between DDA and SWATH MS measurements as a larger difference would impede the use of relative fragment ion intensities as a peptide identification score during SWATH MS data analysis[41].

The optimization of the MS acquisition settings to generate fragment ion spectra for library generation will be described in detail elsewhere (unpublished data). Here we suggest the following generic settings for the acquisition on a TripleTOF 5600: (i) increase fragment ion spectrum accumulation times to 150 ms to maximize the quality of the spectrum; (ii) record more than one fragment ion spectrum of the same precursor by reducing the dynamic exclusion time to 20 s, which is approximately half of a chromatographic peak width (typically

30-60 s). This will increase the chance that a second fragment ion spectrum is recorded from the same sample at higher peptide precursor signal intensity than the first fragment ion spectrum. Further, if the first spectrum is contaminated with fragments of a second, concurrently fragmented precursor, the background would be expected to be changed for the second spectrum. (iii) Aim for the highest similarity possible between the relative intensities of the fragment ions in the library and in the SWATH MS measurements by using the same collision energy settings for both modes of operation. Specifically, regardless of the charge state of the selected peptide precursor, we recommend to use a collision energy which reflects most closely the settings used in SWATH MS data acquisition, for example applying a collision energy according to the equation of a doubly charged peptide (slope 0.0625, intercept -3.5), ramped ±15 V from the calculated collision energy over the MS2 accumulation time (with an upper limit of 80 V). The specific instrument acquisition settings recommended in this protocol are summarized in Supplementary Table 1.

Part 2: Conversion and centroiding of fragment ion spectra

Database search engines that establish peptide-spectrum matches (PSMs) generally require raw instrument output data (profile spectra) to be converted into a peak list format (centroided spectra) in a vendor-independent open format such as mzML[43] or mzXML[44]. Both, the conversion and centroiding (peak picking) process are typically performed by a single tool, the converter. The available centroiding algorithms slightly differ in the way they extract intensities from profile peaks in precursor (MS1) and fragment ion (MS2) spectra. For the purpose of assay library generation, it is important that the converter yields fragment ion intensities that match, as close as possible, those extracted from SWATH MS data. For conversion and centroiding of TripleTOF 5600 fragment ion spectra we tested three different converters: ProteinPilot (AB Sciex), msconvert (with 'prefer vendor' setting; developed by AB Sciex), and qtofpeakpicker, which has been developed by our group and is, like msconvert, also distributed through ProteoWizard (see Supplementary Note 2 for algorithm details). A comparison of the results obtained if sets of DDA files were converted by either of these converters indicates that the qtofpeakpicker (using peak areas) yielded the highest level of reproducibility of fragment ion spectra across replicate DDA runs (Supplementary Figure 1A) and achieved best portability of the derived assays to the corresponding fragment ion intensities obtained by SWATH MS (Supplementary Figure 1B). The numbers of identified peptides and proteins from a database search after conversion with any of the peak pickers are slightly higher for the qtofpeakpicker than those achieved for the other two peak pickers tested (Supplementary Figure 1C). Examples for the converter-dependent variability of relative abundances of fragment ions in centroided MS2 spectra are given in Supplementary Note 4. In summary, these data show that different centroiding algorithms can cause surprisingly large intensity differences for even the most prominent peaks of a fragment ion spectrum. To ensure highest possible assay quality, where accurate relative fragment ion intensities might be crucial for downstream use, a tool that maintains these relative fragment

ion patterns, such as the qtofpeakpicker described above, should thus be selected for the conversion and centroiding of raw instrument files into a search engine-compatible and vendor-independent open format.

Part 3: Sequence database searching and statistical scoring of peptide-spectrum matches

Spectral libraries are built from fragment ion spectra that are assigned with high confidence to a peptide sequence. To establish this match, centroided fragment ion spectra are subjected to sequence database searching. At this stage it is important that the protein sequence database (typically in FASTA format) contains the sequences of the retention time reference peptides to allow for retention time normalization at a later step. To control the false discovery rate (FDR) of the PSMs, the protein sequence database also needs to contain a decoy entry for every protein[45]. Even though protein sequence reversal is, due to its simplicity, the most commonly used method to generate decoy peptides, decoys most precisely reflecting target peptides are generated by pseudo-reversal of target peptide sequences[45]. This latter method was thus used for this protocol. To maximize the number of PSMs and the discrimination between true and false assignments, the search output of multiple search engines may be combined[46]. In general, we recommend using search engines maximally orthogonal in their search algorithms, as this results in highest numbers of identifications[47]. The optimal parameters for search engines, such as number of tolerated tryptic termini, missed cleavages, precursor mass tolerance and variable modifications, depend on the specific biological sample, experimental setup, and purpose of the library. In this protocol, searches were done for fully tryptic peptides. Though semi-tryptic peptides might originate from biologically relevant proteolytic cleavage by endogeneous proteases, several publications have reported that most of those peptides may originate from non-specific trypsin/chymotrypsin activity[48,49] and/or in-source-fragmentation[50]. A discussion on how various parameter settings impact the underlying search has been provided by Eng and colleagues[51]. Notably, the targeted SWATH data analysis does not change upon inclusion or exclusion of semi-tryptic peptides as these will be extracted in the fashion as fully tryptic peptides. Supplementary Table 2 contains the main parameters used for the protocol case study. The PSMs from each search engine are scored using PeptideProphet[52] and subsequently combined and re-scored using iProphet, a tool that integrates evidence from multiple identifications of the same peptide across different experiments and search engines and thus improves discriminating power between correctly and incorrectly assigned PSMs[46].

As dozens or hundreds of DDA runs might be combined to generate a comprehensive SWATH assay library, it is important to thoroughly control the FDR of the final dataset, both at the level of PSMs and the level of inferred proteins. The MAYU software[53] has implemented a robust method to estimate the FDR of such large-scale DDA datasets on PSM, peptide, and protein level and can be applied to the iProphet output. Table 1 shows the FDR at PSM, peptide, and protein level for the dataset associated with this protocol generated from three

DDA files in yeast in comparison to a recently published large scale human library[23] generated from 331 DDA files highlighting the requirement for increased stringency in larger data sets. In some cases the iProphet probability threshold to achieve a certain protein FDR appears low, however, this is unproblematic as the iProphet probability is intended to be interpreted at the peptide level and not at the PSM level. The iProphet probability is used as a ranking and the FDR estimated by MAYU is controlled based on decoys. How the MAYU-estimated FDRs change with the applied iProphet score cut-off is shown in Supplementary Figure 2 for the data set described in this protocol. A discussion of the effect of error rates in spectral libraries for targeted analysis of SWATH/DIA, in particular with respect to error propagation from assay library to SWATH identifications, is provided in **Box 3**.

Part 4: Generation of a spectral library with aligned retention times

SpectraST is a software tool that compiles all fragment ion spectra assigned to a specific peptide sequence above a certain quality threshold (e.g. iProphet probability) into a spectral library format[11,28]. At this step, it is advisable to transform all retention times into a normalized retention time scale[40] before the consensus spectra are computed. This is accomplished by establishing a linear correlation between experimental retention times and unit-less absolute retention time values (iRTs) for retention time reference peptides identified in each DDA run. The resulting correlation curves are then used to convert the retention time for all the other peptides identified in the corresponding DDA runs into iRT scale.

Even though modern mass spectrometers display a reasonably high level of reproducibility in repeat recordings of fragment ion spectra of the same peptide across replicates[41], the consolidation of multiple fragmentation observations of the same peptide precursor ion into a single consensus spectrum provides a more accurate fragmentation pattern than any single spectrum (best replicate) for that precursor[11,54] (Supplementary Figure 3A and B). The consensus spectrum, therefore, is the optimal representation of the fragment ion spectrum of a targeted peptide.

To avoid combining ambiguously assigned fragment ion spectra, e.g. spectra matching to the same sequence but acquired at significantly different normalized retention times into a wrongly averaged consensus spectrum, we developed a strategy to split and process fragment ion spectra from precursors of different retention times for consensus library generation (Figure 2 and **Box 4**). These considerations are particularly important in the context of isobaric peptides with post translational modifications that may be assigned by the search engine to wrong amino acid residues due to ambiguities in the site localization. However, we also observe the phenomenon of distant retention times for identical identifications in datasets which had not been searched for post translational modifications (Supplementary Note 5).

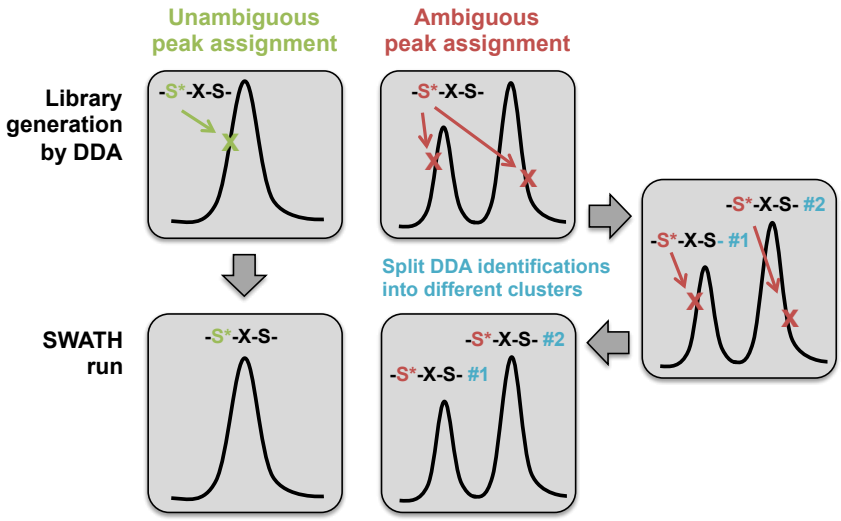

Figure 2 | Splitting peptide identifications with distant elution times.
During a DDA search it may happen that multiple fragment ion spectra are assigned to the same peptide precursor, even though they span a wide retention time segment and might not come from the exact same molecular species. This is not a rare event, especially in the context of post translationally modified peptides where the modification cannot be unambiguously assigned to a certain amino acid. The figure depicts such ambiguous peak assignment on the example of a phospho-peptide containing a phosphorylated serine (S*) in presence of a second, unphosphorylated serine (S). Fragment ion spectra recorded at distant retention times can be clustered apart during the SWATH assay library generation. The distinct SWATH assays might then be used to resolve the correct assignment on the level of SWATH MS data. See Supplementary Note 5 for examples.

Part 5: Generation of a SWATH assay library from a consensus spectral library

Once a consensus spectral library has been generated (or downloaded from the web), the most intense fragment ions need to be retrieved for each precursor. The number of fragment ions should be high enough to ensure specificity of identification within a SWATH MS map, but not too high, as less intense transitions introduce noise into the extracted data, reduce specificity, and may adversely affect the target identification and limit of detection. In previous studies it has been recommended to use six transitions per peptide precursor[9]. Optionally, simulations can be employed to estimate the appropriate number of transitions required to achieve unique ion signatures for the targeted peptides in a given proteome background[55]. Peptide precursors represented in the library with fewer fragment ions than required to achieve high specificity should not be considered. We recommend using the same number of transitions for all assays because different numbers of transitions per precursor may result in mixed statistical distributions for the target identifications in automated peak scoring if this is not accounted for. Regarding ion types, we found that it is acceptable to include y- and b-ions as well as common neutral losses if the library was recorded on the same instrument that is used to record the SWATH MS data[41].

Transitions with fragment ion mass below 350 m/z should be excluded from the library as they are typically less specific and thus more noisy than transitions of fragment ions with higher m/z. Furthermore, also transitions with fragment ion mass falling within the isolation window of their precursor m/z should be excluded from the assay library, as those are

typically highly interfered with incompletely fragmented precursors from the same swath window. Incidentally, this is the only difference in the process of building a library for S/MRM acquisition and for SWATH MS extraction. As this filtering step makes the assay library dependent on the specific instrument setup it is therefore desirable to publish not only the final assay library as a transition list but also the consensus spectral library.

In case a library is to be constructed that contains assays for isotopically light and heavy peptides (e.g. labelled with heavy arginine or lysine at the C-terminus of each peptide), it is important to consider that, depending on the labelling strategy, b-ion transitions from light and heavy precursors might not be distinguishable. This is because these fragments might not carry the isotopic mass difference and the chance is high that the precursor m/z of the light and the heavy form of a peptide fall into the same swath window. For libraries containing C-terminally isotope labelled heavy peptides we thus recommend to only include y-ions in the SWATH assay library.

Different SWATH data analysis software tools accept different formats of assay libraries. The library formats can be divided into two classes: (i) a simple table in tab-separated (tsv) or comma-separated (csv) format where each row contains a transition and columns contain information to specify this transition; (ii) a transition list in TraML format[56]. While the table format is easy to read and manipulate, the TraML format is well defined and thus contains unambiguous information and is the format endorsed by the HUPO Protein Standards Initiative (PSI). OpenSWATH requires libraries in TraML format and containing pre-computed decoy transition groups which facilitate the discrimination between true and false signals and error rate estimation during SWATH data analysis[6]. Decoys need to represent the targets well but at the same time they have to be different from the target assays. Decoys based on shuffled sequences have been shown to be best suited for the purpose of modelling the targets[6]. However, decoys based on full reversal of peptide sequences have been successfully used as well and enable the generation of decoy transition groups for even highly repetitive or palindromic peptide sequences[23]. Other software tools, such as PeakView, Spectronaut, and Skyline require the assay library to be in table format and do not require decoy transition groups to be provided with the library. Supplementary Tables 3 to 6 summarize the assay library formats required for the currently available SWATH analysis software tools. Both, TraML and table format are supported by the tools described in this protocol and the hereby generated SWATH assay libraries can thus be directly used with all major SWATH analysis software suites, namely OpenSWATH, Skyline, Spectronaut, and PeakView.

For the dissemination of SWATH assay libraries, the SWATHAtlas database (www.SWATHAtlas.org) provides a suitable platform. To allow more flexibility for the user, as mentioned above, we advise to publish both, the redundant and consensus spectral libraries in sptxt format, as well as the final assay list in csv/tsv format and/or in TraML format.

Several other databases provide assay/spectral libraries for download, such as SRMAtlas (www.SRMAtlas.org) and PeptideAtlas (www.peptideatlas.org/speclib/), and NIST (http://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:download). However, the libraries provided through these databases might not have been acquired on a TripleTOF mass spectrometer and the relative intensities of fragment ions may thus not reflect the relative intensities of the subsequent SWATH MS measurements very well. Furthermore, it is important to understand whether these libraries contain retention time information at all, if the retention times have been normalized or, most optimally, if they contain iRT values.

MATERIALS

Reagents

- iRT retention time peptides. The retention time peptides are a set of typically 10 to 20 synthetic peptides which span a wide range of hydrophobicity and thus LC retention time which is converted into the unit-free iRT scale[40]. The peptides are spiked into each sample and allow normalizing retention times over various MS runs and different LC setups[40]. Any set of peptides covering a wide range of hydrophobicity can be used. For this protocol we used the iRT-kit from Biognosys.
- One or several peptide samples (See Introduction Part 1A and step 1 of the Procedure for details.)

Equipment

- TripleTOF 5600+ mass spectrometer (AB Sciex)
- Nanoflow HPLC system
- Computer: PC with Microsoft Windows 7 (Microsoft), ≥4 GB of RAM, sufficient hard disk space (for the protocol case study ≥40 GB)
- Software
    - Microsoft Windows 7 (Microsoft)
    - MS Excel (Microsoft)
    - ActivePerl (x86)
    - OpenMS 1.11 - nightly build (32-bit)
    - TPP 4.7 (polar vortex) revision 1
    - Anaconda (32-bit) for Python 2.7 (Continuum Analytics)

Equipment setup

Detailed instructions, including screen shots, for the installation of each software module are provided in Supplementary Note 3.

- **ActivePerl** (x86). Install from http://www.activestate.com/activeperl/downloads with default settings.

- **TPP** 4.7 (polar vortex) rev 1. Install from http://sourceforge.net/projects/sashimi/files/Trans-Proteomic%20Pipeline%20%28TPP%29/ with default settings. (Note: This will also install ProteoWizard's msconvert and qtofpeakpicker, as well as the Apache http server, which is required for the TPP web interface).

  ? TROUBLESHOOTING
- **OpenMS** (version 1.11, 32-bit). For the most updated version (nightly build), install from: http://ftp.mi.fu-berlin.de/OpenMS/nightly_binaries/. OpenMS will ask to install Microsoft .NET Framework 3.5 and 4.0.

  ? TROUBLESHOOTING
- **Anaconda** (32-bit). Install from https://store.continuum.io/cshop/anaconda/ which includes a python interpreter and all required python libraries for Python 2.7. (Select installation for all users to automatically set the location to C:\Anaconda.)
- **msproteomicstools python package**.
  - Open the command line prompt: Open the Start Menu (Windows icon in lower left corner) and type "cmd" in the search field. Right-click the cmd.exe file and select "Run as administrator". To install the package, type:
    ```
    C:\Anaconda\Scripts\pip.exe install msproteomicstools
    ```
  - Alternative ways to install the msproteomicstools package are described in Supplementary Note 3.
- **X!Tandem and Comet parameter files (.params), yeast protein sequence file (yeast.fasta), iRT-kit (iRT.txt), SWATH windows (swaths.txt).** Download from ProteomeXchange[57] (dataset identifier: PXD001126) and move to C:\Inetpub\wwwroot\ISB\data.
- **DDA raw files: nselevse_L120327_001, 010, and 016 (.wiff, .wiff.mtd .wiff.scan).** Download these nine files from ProteomeXchange[57] (dataset identifier: PXD001126) and copy them to C:\Inetpub\wwwroot\ISB\data (the wiff.mtd file can be omitted).
- **SWATH raw files: nselevse_L120412_001, 010, and 016 (.wiff, .wiff.mtd .wiff.scan).** Download these nine files from ProteomeXchange[57] (dataset identifier: PXD001126) and copy them into a folder on your computer.

PROCEDURE

CRITICAL To exemplify the workflow for SWATH library generation, we provide three TripleTOF 5600+ DDA files from a recent SWATH MS study by Selevsek and colleagues (unpublished data). The three files represent samples from an osmotic shock time course (0 min, 60 min, 120 min) in yeast.

All output files from each step of the procedure are listed in Supplementary Note 1 and the parameters of all software tools and commands are described in more detail in Supplementary Note 2.

Part 1: Sample preparation and data acquisition •TIMING Few hours up to several days, depending on the number of samples (excluding preparation of peptide samples)

1| Prepare peptide samples with a final concentration of 0.5-1 µg/µl. How to prepare these peptide samples from biological specimen is highly dependent on the sample type and has been described in many instances in the literature (see for example, human[23] and yeast[58]), however, essentially any sample preparation method compatible with standard shotgun or targeted proteome analysis should be compatible.

2| Spike iRT peptides into your sample at a ratio of 1:20.

CRITICAL STEP. The presence of iRT retention time reference peptides is crucial to generate a high quality SWATH assay library and perform subsequent SWATH MS data analysis. As iRT reference peptides, either a set of synthetic peptides can be used (as described in this protocol, see also the Reagents section), or, alternatively, a set of well detectable endogenous peptides spanning a large retention time range can be used (see main text for more information).
? TROUBLESHOOTING

3| Inject 1-2 µg of your sample onto a nano HPLC coupled to a TripleTOF 5600+ mass spectrometer operating in DDA mode (on the TripleTOF 5600 this is called IDA, information dependent acquisition). Please refer to Supplementary Table 1 for specific instrument parameters.

Part 2: Conversion and centroiding of the raw data •TIMING Few hours, depending on the number of samples

4| Once the data is acquired, copy the files into the folder C:\Inetpub\wwwroot\ISB\data for conversion into a vendor-independent format and centroiding.

5| Open the Start Menu (Windows icon in lower left corner) and type “cmd” in the search field. Right-click the cmd.exe file and select “Run as administrator”.

6| In the command line window that opens, type the following command to change to the directory which contains the data:

```
cd C:\Inetpub\wwwroot\ISB\data
```

! CAUTION Many commands to be entered in the command line window are case-sensitive and it is therefore advised to control the spelling carefully.

7| Run the following command on each file to convert and centroid the profile data:

```
qtofpeakpicker --resolution=20000 --area=1 --threshold=1 --
smoothwidth=1.1 --in nselevse_L120327_001.wiff --out
nselevse_L120327_001.mzXML
```

The --area option causes the converter to use the area of a peak as intensity instead of the peak apex. The converted files will be located in C:\Inetpub\wwwroot\ISB\data. **? TROUBLESHOOTING**

8| Reduce fragment ion spectrum complexity by keeping only the top 150 peaks:

```
msconvert nselevse_L120327_001.mzXML --mzXML --filter
"threshold count 150 most-intense" --outfile
nselevse_L120327_001_c150.mzXML
```

This filtering leads to much smaller file size and, as a consequence, most software tools described in this protocol will run faster. If the library generation workflow is performed using a powerful computing infrastructure this step can be omitted.

Part 3: Database searching and controlling FDR •TIMING Few hours up to several days, depending on the number of samples and search engines to be included

X!Tandem database search

9| To start an X!Tandem search, navigate to the TPP web interface using a web browser, such as Internet Explorer: http://localhost/tpp-bin/tpp_gui.pl (or http://localhost:8080/tpp-bin/tpp_gui.pl). You should also be able to simply double-click the TPP icon that has been generated on your Desktop during the installation.

10| Enter as user name and password "guest" and click on the "Login" button.

11| In the “Home” tab, select as analysis pipeline from the drop-down menu “Tandem”.

12| Go to the "Analysis Pipeline" tab and then to the "Database Search" tab.

13| In the “Specify mz[X]ML Input Files” section, click the "Add Files" button and select the three converted and reduced mzXML files.

14| In the “Specify Tandem Parameter File” section, click the "Add Files" button and select the “xtandem.params” file. This file defines the settings to be used for the search. All settings defined here will overwrite the settings in the default parameter file. Please consult Supplementary Table 2 for a list of parameters which deviate from the default values.

15| In the “Specify a sequence database” section, click the "Add Files" button and select the “yeast.fasta” file. This is a protein sequence database containing all annotated yeast proteins, the iRT retention time reference peptides (concatenated to a single protein) and a pseudo-reversed decoy peptide for each target peptide. The names of the “proteins” containing the decoy peptides start with “reverse_” followed by the target protein name. This tag will be used several times again during the course of this protocol (steps 28|, 29|, 30|, 33|, 35|).

CRITICAL STEP. The sequence database fasta file needs to contain the iRT retention time reference peptides which can be added as separate “proteins” or concatenated to a single protein if the peptides have tryptic ends.

16| Check the option “Convert output to pepXML”.

17| Click the “Run Tandem Search” button.

18| The view switches to the “All Jobs” tab where all jobs which have been submitted recently are listed. Click “refresh” in this table to see the current state. While the X!Tandem search is running the Comet search can already be submitted as well (starting at step 20|).

? TROUBLESHOOTING

19| When the job has finished, you can go to the folder C:\Inetpub\wwwroot\ISB\data and check the three files that were generated for each mzXML file: (i) a .tandem.params file, (ii) a .tandem file, and (iii) a .tandem.pep.xml file.
? TROUBLESHOOTING

Comet database search

20| In the “Home” tab of the TPP web interface, select now as analysis pipeline from the drop-down menu “Comet”.

21| Go to the "Analysis Pipeline" tab and then to the "Database Search" tab.

22| In the “Specify mz[X]ML Input Files” section, click the "Add Files" button and select the three converted and reduced mzXML files (if not selected already from the X!Tandem search).

23| In the “Specify Comet Parameter File” section, click the "Add Files" button and select the “comet.params” file. This file defines the settings to be used for the search. All settings defined here will overwrite the settings in the default parameter file. Please consult Supplementary Table 2 for a discussion of the parameters.

24| In the “Specify a sequence database” section, click the "Add Files" button and select the “yeast.fasta” file (if not selected already). See step 15| for more information on the protein sequence database.

25| Click the “Run Comet Search” button.

26| The view switches to the “All Jobs” tab where all jobs which have been submitted recently are listed. Click “refresh” in this table to see the current state.
? TROUBLESHOOTING

27| When the job has finished, go to the folder C:\Inetpub\wwwroot\ISB\data and check the files that were generated. For each mzXML file, a pep.xml file should have been generated. Add “comet” to the file names of these pep.xml files to avoid confusion later on: xxx.comet.pep.xml (where xxx is the file name).

Score and combine search outputs with PeptideProphet and iProphet. Please note that the following steps can also be done through the TPP web interface (step 28|, 29|, 30|, 33|).

28| To run PeptideProphet on the X!Tandem search results run:

```
xinteract -OARPd -dreverse_ -Ninteract.tandem.pep.xml
nselevse_L120327_0*.tandem.pep.xml
```

29| To run PeptideProphet on the Comet search results run:

```
xinteract -OARPd -dreverse_ -Ninteract.comet.pep.xml
nselevse_L120327_0*.comet.pep.xml
```

30| Run iProphet to combine the search outputs of the X!Tandem and the Comet search and to improve discrimination between true and wrong PSMs:

```
InterProphetParser DECOY=reverse_ interact.comet.pep.xml
interact.tandem.pep.xml iProphet.pep.xml
```

31| To explore which peptides and proteins have been identified by the search engines or inspect the corresponding spectra open the TPP web interface as described in step 9|. Click on the "Utilities" tab, then the “Browse files” tab and select the "iProphet.pep.xml" link in the file list to open the iProphet output in the PepXML viewer.
? TROUBLESHOOTING

32| To export a spread sheet of the iProphet results, click on the “Other Actions” tab and then on the “Export Spreadsheet” button. A file named iProphet.pep.xls is created in the folder C:\Inetpub\wwwroot\ISB\data, which can be opened with Excel.

FDR estimation with MAYU

33| To process the iProphet results with MAYU for FDR estimation, run:

```
Mayu.pl -A iProphet.pep.xml -C yeast.fasta -E reverse_ -G 0.01
-H 51 -I 2 -P protFDR=0.01:t
```

34| Retrieve the minimum iProphet probability at which the protein FDR is <1% by opening the file ending with “_psm_protFDR0.01_t_1.07.csv” in Excel, sorting the column called “score” and reading the lowest value. (For the case study it equals 0.9774, depending on the computer you ran the above software tools the value might be slightly different.)

Part 4: Spectral library generation •TIMING Few hours up to 1 day, depending on the size of the library.

35| To generate a spectral library from all acquired spectra above a certain iProphet cut-off and convert all retention times into iRTs, run the following command after replacing the number following -cP with the cut-off you read out from the MAYU output in the step above:

```
spectrast -cNSpecLib -cICID-QTOF -cf”Protein!~reverse_” -
cP0.9774 -c_IRTiRT.txt -c_IRR iProphet.pep.xml
```

! CAUTION The iRT.txt file contains the peptide sequences to be used as iRT retention time reference peptides. This file needs to be adjusted in case different reference peptides than the ones suggested have been used.

CRITICAL STEP The correlation coefficient R^2 of the linear regression should be > 0.95. Open the spectrast.log file in a text editor and scroll to the end to see the linear regression equation and the R^2.
? TROUBLESHOOTING

36| SpectraST consensus library generation. On high mass accuracy instruments, it may be useful to restrict the merging of spectra for consensus spectrum generation if they have unacceptably large retention time differences. Here we provide two options for consensus library generation. (A) A simple option which assumes that all fragment ion spectra are correctly assigned and (B) a more sophisticated option which additionally considers retention times when merging spectra (see **Box 4**).

A. Consensus library unsplit

i. Generate a consensus library by running the following command:

```
spectrast -cNSpecLib_cons -cICID-QTOF -cAC
SpecLib.splib
```

B. Consensus library split

i. To split distant, retention time-separated peptide identifications, run the following command:

```
python C:\Anaconda\Scripts\spectrast_cluster.py -d
2 SpecLib.sptxt
```

For the case study this command results in 9 output files.

ii. To regenerate .splib, .spidx and .pepidx from the split .sptxt files, run the command:

```
FOR %A IN (SpecLib_*.sptxt) DO spectrast -cNsplit-
%~nA -cICID-QTOF %A
```

(This is equivalent to run 9 times the command: `spectrast -cNsplit-SpecLib_1 -cICID-QTOF SpecLib_1.sptxt` with adjusted numbers.)

iii. Generate a consensus library for each spectral library by running the following command:

```
FOR %A IN (split-SpecLib_*.splib) DO spectrast -
cNcons-%~nA -cICID-QTOF -cAC %A
```

(This is equivalent to run 9 times the command: `spectrast -cNcons-split-SpecLib_1 -cICID-QTOF -cAC split-SpecLib_1.splib` with adjusted numbers.)

iv. Merge the consensus libraries back into a single consensus library:

```
grep -hUv ### cons-split-SpecLib_*.sptxt >>
SpecLib_cons_concat.sptxt
```

(grep is a little executable which is installed together with the TPP.)

The splitting will add the tag “Subgroup_xx_” in front of the protein name so that the different clusters of a peptide can be identified easily.

Part 5: Assay library generation •TIMING 30 min up to few hours, depending on the size of the library

37| The last step is to convert a spectral library into an assay library for SWATH MS data analysis, i.e. to extract the most intense fragment ions for each peptide precursor. The SWATH windows can be defined in a simple table, which allows the script to disregard transitions for which the fragment ion falls into the same window as the precursor ion, as these typically result in noisy signals. If you plan to analyse your SWATH data with PeakView follow option A, for Skyline, Spectronaut, or OpenSWATH follow option B. The required input formats for each SWATH analysis software are listed in Supplementary Tables 3 to 6.

A. For PeakView

i. Extract most intense transitions from spectral library:

```
python C:\Anaconda\Scripts\spectrast2tsv.py -l 350,2000 -s b,y -x 1,2 -o 6 -n 6 -p 0.05 -d -e -w swaths.txt -k peakview -a SpecLib_cons_peakview.tsv SpecLib_cons.sptxt
```

To run this command for the split library, replace the input file at the end of the command to "SpecLib_cons_concat.sptxt".
The swaths.txt file contains the swath windows which are required to ignore transitions with fragment ion m/z falling into their precursor swath window.
! CAUTION The swaths.txt file needs to be adjusted to contain the SWATH window scheme that has been (or is to be) used for SWATH data acquisition.
The spectrast2tsv.py script recognizes common amino acid modifications, but if required, additional ones can be specified using an additional input table. An example for this can be found in the msproteomicstools folder under analysis\spectral_libs\config_file_examples.

B. For OpenSWATH, Spectronaut and Skyline

i. Extract most intense transitions from spectral library to generate the final SWATH assay library:

```
python C:\Anaconda\Scripts\spectrast2tsv.py -l 350,2000 -s b,y -x 1,2 -o 6 -n 6 -p 0.05 -d -e -w swaths.txt -k openswath -a SpecLib_cons_openswath.csv SpecLib_cons.sptxt
```

To run this command for the split library, replace the input file at the end of the command to "SpecLib_cons_concat.sptxt".
The swaths.txt file contains the swath windows which are required to ignore transitions with fragment ion m/z falling into their precursor swath window.
! CAUTION The swaths.txt file needs to be adjusted to contain the SWATH window scheme that has been (or is to be) used for SWATH data acquisition.
The spectrast2tsv.py script recognizes common amino acid modifications, but if required, additional ones can be specified using an additional input table. An example for this can be found in the msproteomicstools folder under analysis\spectral_libs\config_file_examples.

ii. The OpenSWATH software requires the SWATH assay library to be in TraML format:

```
ConvertTSVToTraML -in SpecLib_cons_openswath.csv -out SpecLib_cons.TraML
```

iii. The OpenSWATH software requires decoy transition groups to be present in the TraML assay library. Add decoy transition groups based on shuffled sequences:

```
OpenSwathDecoyGenerator -in SpecLib_cons.TraML -out SpecLib_cons_decoy.TraML -method shuffle -append -exclude_similar
```

In the Supplementary Tutorial we describe how to load the library exported in step 37| B(i) into Skyline and how to extract SWATH traces for visualization and data analysis.

BOXES

Box 1 | Terminology

Centroid / profile mode	Raw data (profile mode) is peak-picked (centroided) to produce a peak list of precursor ion masses (MS1) and of fragment ion masses (MS2) which can be used in peptide identification by database searching.
Data-dependent acquisition (DDA)	Mode of operation of a tandem mass spectrometer in which a fixed number of the most abundant precursor ions (e.g. top 20) in every MS1 survey scan are selected for fragmentation and subsequent recording of an MS2 scan. This strategy is commonly referred to as 'shotgun proteomics'.
Data-independent acquisition (DIA)	Mode of operation of a tandem mass spectrometer which uses a fixed duty cycle to acquire tandem mass spectra from mixed populations of precursor ions which have been co-fragmented using isolation windows ranging from tens to hundreds of m/z units. Comprehensive MS2 spectral coverage over a large mass range can be achieved by iterating over sequential precursor isolation windows in a single duty cycle.
Decoy	Additional peptide/protein sequences concatenated to the main protein sequence database which are used to estimate the false discovery rate in database searching. The decoys should be representative of the target proteins in number and composition, are typically generated by pseudo-reversal, reversal, or scrambling of the target protein sequences and should not be contained in the searched database.
False discovery rate (FDR)	An estimate of the number of false positive identifications contained in a database search result at a given score threshold. FDR can be estimated at the PSM, peptide, and protein levels.
Peptide-spectrum match (PSM)	A confident assignment by a database search engine of a peptide sequence to a single MS2 spectrum acquired in DDA mode.
Selected/Multiple reaction monitoring (S/MRM)	Mode of operation of a triple quadrupole mass spectrometer in which the first quadrupole is fixed on the precursor m/z of a given peptide, the precursor is fragmented in the collision cell, and the third quadrupole is fixed on a fragment ion. The instrument cycles through a fixed list of Q1/Q3 pairs (transitions – see below) and intensities are recorded over chromatographic time. Considered the gold standard for peptide quantification.
Transition	A pair of masses that represent the precursor ion and a single fragment ion from a given peptide. Multiple transitions are measured in an S/MRM experiment to unambiguously identify and quantify a peptide.
SWATH MS	An instance of the DIA strategy in which highly multiplexed MS2 spectra are collected from wide precursor windows which are designed to cover the m/z range expected for tryptic peptides in a cycle time that is short compared to the elution time of a peptide (e.g. 32 windows of 25 m/z width acquired at a dwell time of 100 ms per window). Quantitative data is extracted in a targeted fashion based on prior knowledge of mass spectrometric and chromatographic behaviour of peptides using an assay library.
Indexed retention time (iRT)	A normalized retention time space calibrated using synthetic peptides which are spiked into every sample measured.
Spectral library	A collection of MS2 spectra with high confidence peptide sequence assignments.
Consensus spectral library	A spectral library in which MS2 spectrum entries with a redundant peptide sequence assignment have been collapsed into a single entry.
Assay library	A set of coordinates used for targeted extraction of SWATH/DIA data which typically includes the peptide sequence, the precursor m/z and charge state, the most intense fragment ions m/z and charge states, relative fragment ion intensities, and iRT.

Box 2 | Library generation check list

This box is meant to summarize critical considerations to be made during SWATH assay library generation.

DDA data acquisition

☐ The samples might be pre-fractionated (e.g. by OGE) to increase coverage.

☐ The samples contain reference peptides for retention time normalization.

☐ The DDA instrument parameters are optimized for high-quality fragment ion spectrum acquisition (i.e. longer acquisition/dwell time/trap filling, shorter dynamic exclusion).

☐ The DDA collision energy, including ramping, mimics the one to be used to fragment that same precursor in SWATH acquisition.

☐ The spectra in DDA files are centroided with a suitable converter, optimally using fragment ion peak areas instead of peak height for centroiding.

☐ The DDA files are converted to centroid mode without de-isotoping.

DDA database search and spectral library generation

☐ When multiple search engines are to be used, the DDA data files are converted to the various input formats with consistent spectrum indices.

☐ The multiple search engine results are aggregated using adequate tools (e.g. iProphet).

☐ The protein FDR of the raw spectral library is tightly controlled to be ≤1%.

☐ The retention times in the raw spectral library are aligned to reference values (e.g. iRT) before the consensus library generation.

☐ Optional (mainly recommended for assay libraries with post translational modifications): Precursors in the raw spectral library are split into as many clusters as needed based on their normalized retention time before consensus library generation.

Assay library / transition list generation

☐ Fragments smaller than 350 m/z or bigger than 2000 m/z are filtered out.

☐ Fragments with m/z in the precursor swath window are filtered out.

☐ Only fragments with mass accuracy within ±0.05 m/z of the expected mass are used.

☐ The most intense y and b-ion fragments fulfilling the above criteria are selected.

☐ In the case of a library containing assays for C-terminally heavy isotope-labelled peptides, no b-ions must be included.

☐ Fragments with neutral loss may be considered if the library was acquired on the same instrument.

☐ All assays should have the same number of fragment ions.

Box 3 | Considerations for controlling the spectral library false discovery rate

Estimating and controlling the FDR (false discovery rate) in shotgun proteomics has been the subject of many studies and standard methods using model-based[52,59] and decoy-based[45] approaches, or hybrids of these, are now well established in the field. As the scale of proteomics projects has grown, and the scanning speed of mass spectrometers has increased, it has become apparent that methods which deal specifically with robustly estimating FDRs in very large-scale datasets are required. Such methods have been developed and implemented in the MAYU software[53] with a particular focus on estimating the PSM (peptide spectrum match) level, peptide level, and protein level FDR in large-scale DDA datasets. As the creation of very large spectral libraries for use in targeted SWATH data analysis workflows is actively being pursued, a discussion of the effect of error rates in spectral libraries built for targeted analysis of SWATH/DIA is justified.

Previous studies have emphasized the importance of high quality spectra when constructing spectral libraries and suggested that errors introduced at this stage might be propagated into the results of a spectral library searching strategy for DDA data[11]. A question which has not been directly addressed is whether errors in spectral libraries will be propagated into targeted analysis of SWATH data, or whether the consistent fragment ion spectrum sampling in chromatographic time will be able to resolve errors which are introduced at the spectral library level. This question can only be answered by considering the source of error in the spectral library. For example, if the error arises because of co-isolation of multiple peptide species (or other species) then a mixed (or chimeric) spectrum will result with the potential to match to a peptide in the sequence database with a high score. If such a library spectrum is then used as the basis for targeted analysis of SWATH data, it is improbable that a high score will be produced because the fragment ions are very unlikely to perfectly coelute and, as such, the error from the spectral library will not be propagated to the SWATH data analysis results.

However, there is a second type of spectral library error which is more problematic. That is, a fragment ion spectrum in the DDA data could be produced from a single peptide precursor and still match to the wrong sequence in the database search. If this is the case, SWATH data, and targeted analysis thereof, will match faithfully to the library spectrum with perfect co-elution of fragment ions, thereby propagating the original error into the SWATH analysis results. If the first type of error is predominant in the DDA data then a moderate FDR in the spectral library would be well tolerated for downstream analysis and propagation to SWATH results would not be an issue. However, if the second type of error is more frequent, a more conservative FDR threshold in the library creation would be required. To our knowledge a systematic investigation into which type of error predominates in DDA data has not been performed but remains an open question worthy of further study. With this uncertainty researchers may choose the threshold depending on the downstream analysis question, but perhaps for large-scale libraries which could be distributed for use by many labs a more conservative threshold is warranted. In any case, robust methods for estimating FDR at PSM, peptide, and protein level, such as provided by MAYU, should be employed during the library creation process.

Box 4 | Splitting peptide identifications with distant elution times

On a high performance liquid chromatography system, any given peptide is expected to elute within a single peak at a characteristic, well-defined retention time. In a DDA workflow it may happen, however, that multiple identifications of a given peptide actually span a rather large time segment, eventually longer than that covering the average peptide chromatographic peak width within the HPLC condition used. In such cases, questions may arise for consensus library generation whether those multiple fragment ion spectra should be globally combined into a single assay or whether they should be clustered apart and processed independently. Despite having the same peptide identification, those fragment ion spectra

could indeed originate from different isobaric peptide sequences (falsely assigned) or different peptide conformations, and therefore a global consensus merging would erroneously combine those different fragmentation spectra, yielding a single peptide assay with both, incorrect relative intensities and incorrect retention time approximation (Figure 2). For a best replicate-based non-redundant spectral library, intensity and retention time would be correctly retrieved for one spectrum cluster but all information on the other spectrum cluster(s) would be lost.

This phenomenon is quite frequently observed with peptides carrying post translational modifications where it can be challenging for a search engine to accurately locate it on the peptide sequence (unpublished data). In this library generation protocol we present an optional extended subroutine to handle such cases. The additional steps consist of (i) the separation of clusters of fragment ion spectra with elution time beyond a user-defined threshold (e.g. two iRT units for our HPLC setup), (ii) the generation of consensus spectra for each of these clusters independently, and (iii) the merging of all those consensus spectra into a final consensus spectral library. The resulting assay library thus contains multiple assays (with different relative intensities and retention times) associated to the same peptide sequence. The uniqueness of the peptide assays is provided either by using a unique assay identification number (openswath option in step |37) or by using an incremented protein name (peakview option in step |37) for each assay pointing to the same peptide sequence. By using this pipeline, several compelling cases of peptide identifications were found in the provided datasets even without searching for post translational modifications (Supplementary Note 5).
It should be noted that the quantification of those multiple chromatographic peaks matching to identical peptide sequences is not trivial and it is beyond the scope of this protocol to describe the detailed downstream analysis steps for such cases.

TIMING

Steps 1-3 (part 1), few hours up to several days, depending on the number of samples (excluding preparation of peptide samples)
Steps 4-8 (part 2), few hours, depending on the number of samples
Steps 9-34 (part 3), few hours up to several days, depending on the number of samples and search engines to be included
Steps 35-36 (part 4), few hours up to 1 day, depending on the size of the library
Steps 37 (part 5), 30 min up to few hours, depending on the size of the library
For the specific data set described in this protocol we required three days for sample preparation (part 1, including sample preparation) and two days for the bioinformatic part (parts 2-5).

TROUBLESHOOTING

Troubleshooting advice can be found in Table 2.

ANTICIPATED RESULTS

The final SWATH assay library generated from the three example yeast injections consists of 101,472 transitions belonging to 16,912 peptide precursors, 15,239 modified peptides, 14,948 stripped peptides, and 1948 unique proteins and can be directly used as an input for all currently available SWATH analysis software tools (the numbers correspond to the SWATH assay library without RT-based splitting).

The DDA and SWATH raw files of the case study have been deposited to the ProteomeXchange Consortium[57] via the PRIDE partner repository with the dataset identifier PXD001126.

ACKNOWLEDGEMENTS

We would like to thank Christina Ludwig and Samuel Bader for discussions and feedback on the manuscript, Lorenz Blum for implementation of iProphet support in MAYU, Hannes Röst for packaging msproteomicstools, Joe Slagel for including the qtofpeakpicker and the new MAYU version in the TPP, and the PRIDE Team for maintaining the ProteomeXchange platform. This work has been financially supported by the Framework Programme 7 of the European Commission through SysteMTb (241587), UNICELLSYS (201142), PRIME-XS (262067), and ProteomeXchange (260558), an ERC advanced grant ''Proteomics v3.0'' (233226), the Federal Ministry of Education and Research (e:Bio Express2Present, 0316179C), and the Forschungszentrum Immunologie of the University Medical Center Mainz.

AUTHOR CONTRIBUTIONS

OS, LG, and BC developed the workflow and wrote the manuscript; PN developed the tools spectrast2tsv.py and spectrast_cluster.py; GR and HL developed and implemented the retention time normalisation and iRT calibration in SpectraST; WW developed the qtofpeakpicker; DA, PM and BM implemented automated SWATH library import into Skyline; RA directed the project and contributed to writing the manuscript.

COMPETING FINANCIAL INTERESTS

R.A. holds shares of Biognosys AG, which operates in the field covered by the article (products are Spectronaut software and iRT-kit). The authors have declared no conflict of interest.

REFERENCES

1. Domon, B. & Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. *Nature Biotechnology* **28,** 710–721 (2010).
2. Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nature Methods* **9,** 555–566 (2012).
3. Gillet, L. C. *et al.* 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,** O111.016717 (2012).
4. Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. & Yates, J. R., III. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods* **1,** 39–45 (2004).
5. Chapman, J. D., Goodlett, D. R. & Masselon, C. D. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrom Rev* (2013). doi:10.1002/mas.21400
6. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology* **32,** 219–223 (2014).
7. Bernhardt, O. M. *et al.* Spectronaut: A fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data. in *Proceedings 60th ASMS Conference on Mass Spectrometry* 1–2 (2012).
8. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics (Oxford, England)* **26,** 966–968 (2010).
9. Reiter, L. *et al.* mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nature Methods* **8,** 430–435 (2011).
10. Zi, J. *et al.* Expansion of the ion library for mining SWATH-MS data through fractionation proteomics. *Anal Chem* **86,** 7242–7246 (2014).
11. Lam, H. *et al.* Building consensus spectral libraries for peptide identification in proteomics. *Nature Methods* **5,** 873–875 (2008).

12. Hughes, M. A., Silva, J. C., Geromanos, S. J. & Townsend, C. A. Quantitative proteomic analysis of drug-induced changes in mycobacteria. *J Proteome Res* **5,** 54–63 (2006).
13. Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S. & MacCoss, M. J. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal Chem* **78,** 5678–5684 (2006).
14. Picotti, P. *et al.* A database of mass spectrometric assays for the yeast proteome. *Nature Methods* **5,** 913–914 (2008).
15. Prakash, A. *et al.* Expediting the Development of Targeted SRM Assays: Using Data from Shotgun Proteomics to Automate Method Development. *J Proteome Res* **8,** 2733–2739 (2009).
16. Picotti, P. *et al.* High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nature Methods* **7,** 43–46 (2010).
17. Collins, B. C. *et al.* Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nature Methods* (2013). doi:10.1038/nmeth.2703
18. Picotti, P. *et al.* A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* **494,** 266–270 (2013).
19. Schubert, O. T. *et al.* The Mtb Proteome Library: A Resource of Assays to Quantify the Complete Proteome of Mycobacterium tuberculosis. *Cell Host Microbe* **13,** 602–612 (2013).
20. Karlsson, C., Malmström, L., Aebersold, R. & Malmström, J. A. Proteome-wide selected reaction monitoring assays for the human pathogen Streptococcus pyogenes. *Nature Communications* **3,** 1301 (2012).
21. Hüttenhain, R. *et al.* N-Glycoprotein SRMAtlas: a resource of mass-spectrometric assays for N-glycosites enabling consistent and multiplexed protein quantification for clinical applications. *Mol Cell Proteomics* (2013). doi:10.1074/mcp.O112.026617
22. Hüttenhain, R. *et al.* Reproducible quantification of cancer-associated proteins in body fluids using targeted proteomics. *Sci Transl Med* **4,** 142ra94 (2012).
23. Rosenberger, G. *et al.* A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data* **1,** 140031 (2014).
24. Deutsch, E. W. *et al.* A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **10,** 1150–1159 (2010).
25. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology* **30,** 918–920 (2012).
26. Sturm, M. *et al.* OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics* **9,** 163 (2008).
27. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26,** 1367–1372 (2008).
28. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7,** 655–667 (2007).
29. Lam, H. & Aebersold, R. Building and searching tandem mass (MS/MS) spectral libraries for peptide identification in proteomics. *Methods* **54,** 424–431 (2011).
30. Weisbrod, C. R., Eng, J. K., Hoopmann, M. R., Baker, T. & Bruce, J. E. Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification. *J Proteome Res* **11,** 1621–1632 (2012).
31. Heller, M. *et al.* Added value for tandem mass spectrometry shotgun proteomics data validation through isoelectric focusing of peptides. *J Proteome Res* **4,** 2273–2282 (2005).
32. Stergachis, A. B., MacLean, B., Lee, K., Stamatoyannopoulos, J. A. & MacCoss, M. J. Rapid empirical discovery of optimal peptides for targeted proteomics. *Nature Methods* **8,** 1041–1043 (2011).
33. Qeli, E. *et al.* Improved prediction of peptide detectability for targeted proteomics using a rank-based algorithm and organism-specific data. *Proteomics* (2014). doi:10.1016/j.jprot.2014.05.011
34. Mallick, P. *et al.* Computational prediction of proteotypic peptides for quantitative proteomics. *Nature Biotechnology* **25,** 125–131 (2006).
35. Eyers, C. E. *et al.* CONSeQuence: prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches. *Mol Cell*

Proteomics **10,** M110.003384 (2011).
36. Fusaro, V. A., Mani, D. R., Mesirov, J. P. & Carr, S. A. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nature Biotechnology* **27,** 190–198 (2009).
37. Tang, H. *et al.* A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics (Oxford, England)* **22,** e481–8 (2006).
38. Webb-Robertson, B.-J. M. *et al.* A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. *Bioinformatics (Oxford, England)* **24,** 1503–1509 (2008).
39. Li, S., Arnold, R. J., Tang, H. & Radivojac, P. On the Accuracy and Limits of Peptide Fragmentation Spectrum Prediction. *Anal Chem* **83,** 790–796 (2011).
40. Escher, C. *et al.* Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **12,** 1111–1121 (2012).
41. Toprak, U. H. *et al.* Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a discriminating feature for targeted proteomics. *Mol Cell Proteomics* (2014). doi:10.1074/mcp.O113.036475
42. de Graaf, E. L., Altelaar, A. F. M., van Breukelen, B., Mohammed, S. & Heck, A. J. R. Improving SRM assay development: a global comparison between triple quadrupole, ion trap, and higher energy CID peptide fragmentation spectra. *J Proteome Res* **10,** 4334–4341 (2011).
43. Deutsch, E. mzML: a single, unifying data format for mass spectrometer output. *Proteomics* **8,** 2776–2777 (2008).
44. Keller, A., Eng, J., Zhang, N., Li, X.-J. & Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Molecular Systems Biology* **1,** 2005.0017 (2005).
45. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4,** 207–214 (2007).
46. Shteynberg, D. *et al.* iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol Cell Proteomics* **10,** M111.007690 (2011).
47. Shteynberg, D., Nesvizhskii, A. I., Moritz, R. L. & Deutsch, E. W. Combining Results of Multiple Search Engines in Proteomics. *Mol Cell Proteomics* (2013). doi:10.1074/mcp.R113.027797
48. Picotti, P., Aebersold, R. & Domon, B. The implications of proteolytic background for shotgun proteomics. *Mol Cell Proteomics* **6,** 1589–1598 (2007).
49. Walmsley, S. J. *et al.* Comprehensive analysis of protein digestion using six trypsins reveals the origin of trypsin as a significant source of variability in proteomics. *J Proteome Res* **12,** 5666–5680 (2013).
50. Kim, J.-S., Monroe, M. E., Camp, D. G., Smith, R. D. & Qian, W.-J. In-source fragmentation and the sources of partially tryptic peptides in shotgun proteomics. *J Proteome Res* **12,** 910–916 (2013).
51. Eng, J. K., Searle, B. C., Clauser, K. R. & Tabb, D. L. A face in the crowd: recognizing peptides through database search. *Mol Cell Proteomics* **10,** R111.009522 (2011).
52. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **74,** 5383–5392 (2002).
53. Reiter, L. *et al.* Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol Cell Proteomics* **8,** 2405–2417 (2009).
54. Liu, J. *et al.* Methods for peptide identification by spectral comparison. *Proteome science* **5,** 3 (2007).
55. Röst, H. L., Malmström, L. & Aebersold, R. A computational tool to detect and avoid redundancy in selected reaction monitoring. *Mol Cell Proteomics* **11,** 540–549 (2012).
56. Deutsch, E. W. *et al.* TraML--a standard format for exchange of selected reaction monitoring transition lists. *Mol Cell Proteomics* **11,** R111.015040 (2012).
57. Vizcaíno, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology* **32,** 223–226 (2014).
58. Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B. & Aebersold, R. Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **138,** 795–806

(2009).
59. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Anal Chem* **75,** 4646–4658 (2003).

TABLES

Table 1| Effect of increasing data set size on score threshold and false discovery rate.

iProphet probability threshold	PSM FDR	Peptide FDR	Protein FDR	Protein identifications	Dataset
0.9774	0.08 %	0.20 %	**1.0 %**	2,162	Yeast 3 DDA files (this protocol)
0.9171	0.16 %	0.40 %	**2.0 %**	2,249	
0.3983	0.37 %	1.04 %	**5.0 %**	2,414	
0.9994	0.07 %	0.18 %	**1.0 %**	11,102	Human 331 DDA files (Rosenberger *et al.*)
0.9970	0.19 %	0.38 %	**2.0 %**	11,537	
0.9809	0.44 %	1.00 %	**5.0 %**	12,203	

Table 2 | Troubleshooting table.

Step (where observed, not where occurred)	Problem	Possible reason	Solution
2, 39	iRT peptides not spiked into sample or R^2 of iRT calibration < 0.95	iRT peptides not present in sample or not all iRT peptides were correctly identified / detectable	In case no iRT peptides were spiked into the samples or if they are not well detectable (or they show a bad correlation), it is possible to use endogenous peptides present in all samples to perform the retention time alignment as described by Parker et al., submitted. Alternatively, a higher concentration of the iRT peptides could be spiked into the samples to make sure they can be correctly identified.
7	Error when running command and during installation of OpenMS/ProteoWizard	A software dependency for the ProteoWizard tools (qtofpeakpicker and/or msconvert) is not present	ProteoWizard requires .NET Framework 3.5 SP1 and 4.0 and Microsoft Visual C++ 2008 SP1 and 2010 SP1 Redistributable Packages to be installed. These should have been installed automatically during the installation of OpenMS, however, in case the step was skipped or resulted in an error then these packages might need to be installed manually. Go to Control Panel → Programs and Features → Turn Windows features on or off → activate Microsoft .NET Framework 3.5.1 (including both subfolders). In our experience the error can even be ignored.
7, 19, 28	Error when running command	User does not have permissions to write	Right click on the folder and select “Properties”. Go to the “Security” tab and

		a file into the target folder	allow the users “Full control”.
20	pepXML not generated for all X!Tandem search outputs	Bug in TPP web interface	Go to the “Analysis Pipeline (Tandem)” tab and then to the “pepXML” tab. Select here the .tandem files which have not been converted to pepXML and click the “Convert to PepXML” button.
28	Comet gives an error	TPP version 4.7.0 comes with older comet version which is not compatible with current parameter file	Uninstall TPP and install version 4.7.1 as described in the installation part of the protocol.
19, 28	Comet or X!Tandem gives an error	Search takes too long and Apache reaches timeout.	Click on the Windows button in the lower left corner and in the search filed enter httpd.conf. It should suggest “Edit the Apache httpd.conf Configuration File”. Scroll down until you see close to the end of the file the lines: # Add 5-hour timeout Timeout 18000 Replace these two lines with: # Add 5-hour timeout - changed to 1 week Timeout 604800 Save and close. Your user needs full rights on that file to be allowed to modify it: right click on the file and select “Properties”. Go to the “Security” tab and allow the users “Full control”.
33, 34	Protein of interest not covered in IDA runs for library generation	Protein too low abundant or not expressed under the condition used for library generation	Libraries which do not cover the proteins of interest could be topped up with synthetic peptides. Crude unpurified synthetic peptides can be purchased from JPT or Thermo for 10-20 EUR/USD per peptide.
To get further support for ProteoWizard tools, please consult the website (http://proteowizard.sourceforge.net) and email to the support list (support@proteowizard.org).			
To get further support for TPP tools, please consult the Wiki (http://tools.proteomecenter.org/wiki/index.php?title=Software:TPP) and subscribe to the spctools discussion group (https://groups.google.com/forum/#!forum/spctools-discuss).			
To get further support for OpenMS tools, please consult the website (http://open-ms.sourceforge.net) and subscribe to the support mailing list (http://sourceforge.net/p/open-ms/mailman).			