

Characterization of QC DDA experiments

David L. Tabb, August 18, 2018

Contents

Introduction	1
Manual evaluation for QC data.....	2
Assessing the end product: quantifying identification success	3
Accelerating chromatogram examination through Skyline software	7
References	9

Introduction

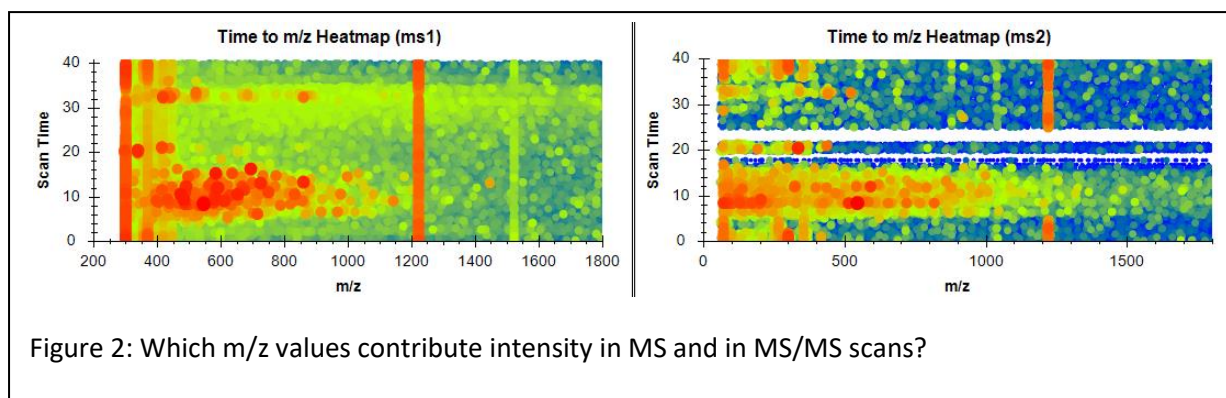
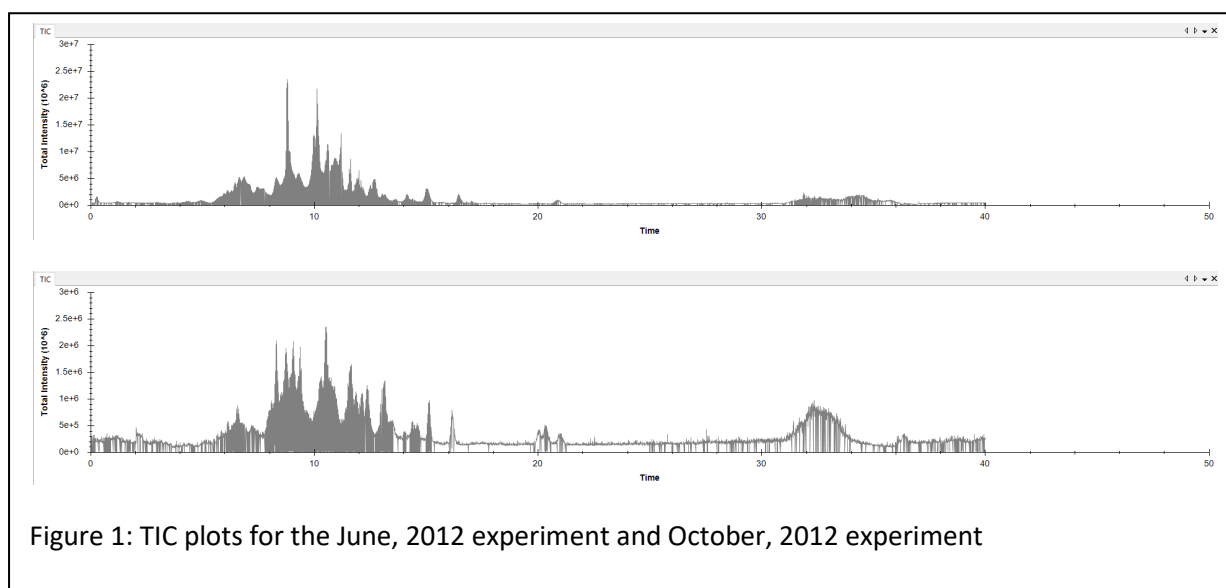
Many proteomics laboratories test their instruments through the repeated analysis of defined protein mixtures. One of the simplest routes produces an LC-MS/MS experiment from the digest of an individual protein, such as bovine serum albumin. Other opt for a more complex mixture, such as the Sigma-Aldrich “Universal Proteomics Standard 1,” which comprises an equimolar mixture of 48 human proteins, or the Agilent “Complex Proteomics Standard,” which comprises a whole-cell lysate of *Pyrococcus furiosus*. Of course, the lysate of *Saccharomyces cerevisiae* continues to be a popular option for many laboratories¹. Any of these mixtures will contain a great diversity of peptides to test separation and mass analysis of a tandem mass spectrometer.

The Association for Biomolecular Resource Facilities (ABRF) conducted a nine-month study among 64 core laboratories around the world to evaluate the longitudinal reproducibility of tandem mass spectrometry². ABRF opted to use a six-protein bovine mixture that had been predigested by Michrom Bioresources. Each lab was asked to produce an LC-MS/MS experiment from the samples once a month over the period of the study. In this tutorial, we will examine eight LC-MS/MS files produced by participant 364386 (participant identities are anonymized) using an Agilent QTOF, model 6530. The data have been transformed from the raw format, directories with the “.d” extension, to the mzML format for easier management via the ProteoWizard msConvert tool using the “peakPicking” filter³.

The most common way for laboratories to evaluate their QC runs rely strongly on manual interpretation. We will start with this approach. Another common strategy requires that we produce identifications from the LC-MS/MS data; we will compare some strategies for interpreting those identifications. Finally, we will employ the Skyline software (most typically associated with Selected Reaction Monitoring experiments) to automate several steps and support the comparison of multiple replicate injections.

Manual evaluation for QC data

Some of the most common variables to be tested in QC data are signal intensity and properties of individual peptides. Researchers often have a target TIC (Total Ion Current) or BPI (Base Peak Intensity) in mind for each experiment. In practice experiments, may vary considerably from this target. Figure 1 shows two TIC plots drawn from the set of eight experiments; these were created through the “SeeMS” software in ProteoWizard. Even though these experiments were intended to be run as identically as possible, we see that the y-axis maximum for June is 30,000,000 while the maximum for October is only 3,000,000. The intensity difference may be the reason that the late-eluting ions (perhaps incompletely digested proteins) appear to be a more significant contribution to intensity.



The SeeMS software offers another plot, the Time to m/z Heatmap, which may be of use in evaluating QC runs. The TIC plot is indifferent to which ions are contributing intensity at any given point in retention time. The heatmap (Figure 2) reveals which m/z values contribute intensity across the LC gradient. Polymer contaminants such as PEG may show ions that repeat at patterns of 44 m/z across these heatmaps. Co-electrosprayed calibrants should also appear as reliable signals across the retention time window.

Since peptide ions are the ones we most hope to see in these QC experiments, it stands to reason that most researchers seek out favorite peptides to check mass spectrometer performance (of

course, if samples are digested on-site by the same protocol as experimental samples, one can also check the performance of digestion). The Michrom mixture contains beta lactoglobulin, lactoperoxidase, carbonic anhydrase, glutamate dehydrogenase, alpha casein, and serum albumin. The digestion used is somewhat unusual in that iodoacetic acid was used to modify free sulfhydryls, so Cys residues have a mass raised by 58 Da rather than the 57 Da associated with iodoacetamide. In this case, we will look for one dominant peptide from each of the six proteins in the June, 2012 experiment:

LGEYGFQNALIVR (+2), LSFNPTQLEEQC[58]HI (+2), YLGYLEQLLR (+2), DDGSWEVIEGYR (+2), VGPLLAC[58]LLGR (+2), and AVVQDPALKPLAL (+2). For convenience, we will refer to each by its first three amino acids.

LGE in its +2 charge state appears at an m/z of 740.4013 (produced from the PeptideMass tool in ExPASy⁴: <http://web.expasy.org>). From experience, we know that its retention time in this chromatography places it at approximately 12 minutes. In the June, 2012 experiment, it first stands clear of the noise at scan 715224, 11.92 minutes into the separation. It continues to scan 723586 at 12.06 minutes. Its m/z is quite steady during this interval, with SeeMS reporting either 740.40 or 740.41 m/z (I did not remember how to coerce the user interface to show three decimal places). At no time did this ion become the base peak.

LSF (+2) should appear at 858.8985 (remember to include +58 for Cys). We expect it to appear at 12.6 minutes. In the June, 2012 experiment, we see it appearing at 12.546 min, and by 12.676 min it serves as the base peak (the tallest in the mass spectrum). Around 12.912 min, however, something odd happens. The monoisotope peak falls away and a new isotopic packet rises, eventually becoming a base peak at 859.39. We are reminded that MS scans are very “busy,” and without very high resolution, we may easily confuse one peptide chromatogram with another.

We may work through each of the remaining peptides. This table of m/zs and identified scan numbers should help:

Sequence	m/z	Ret. Time
LGEYGFQNALIVR (+2)	740.4013	11.920
LSFNPTQLEEQC[58]HI (+2)	858.8985	12.676
YLGYLEQLLR (+2)	634.3559	14.839
DDGSWEVIEGYR (+2)	713.3177	10.394
VGPLLAC[58]LLGR (+2)	585.3392	15.329
AVVQDPALKPLAL (+2)	667.9057	11.939

Performing this examination of chromatograms can seem taxing for someone first assessing a laboratory standard experiment. Over time, of course, the operators gain familiarity with the software and may even memorize m/z and retention times.

Assessing the end product: quantifying identification success

When the goal of a core laboratory’s workflow is the identification of proteins, it makes sense that the facility technicians would monitor the performance of quality control experiments by looking at the identifications produced from these files. While the manual analysis from the prior section can be performed on raw data in minutes after the QC experiment is performed, identification

assessment requires more extensive computation before evaluation can begin. Generally the computation includes three steps⁵:

1. Converting the raw data to a text format the search engine can read (e.g. mzML, MGF)
2. Comparing MS/MS scans to predicted spectra generated from FASTA peptides
3. Filtering the Peptide-Spectrum Matches to a target FDR and inferring proteins

This slow step can be mitigated along several lines. First, some search engines can read raw files directly (e.g. Spectrum Mill, MaxQuant⁶, and MyriMatch⁷). Second, one can use a much smaller sequence database that is tuned to the results from past QC experiments (though this can foul the down-stream statistics, in some cases). Third, a lab could use spectral library search for QC data; this is a near-ideal use case for systems like SpectraST⁸ and Bibliospec⁹.

In this case, these raw data were converted to mzML peaklists in ProteoWizard msConvert:

```
msconvert.exe --filter "peakPicking true 1-" -z *.d
```

The mzMLs were searched with the MS-GF+ database search engine¹⁰, using the Ensembl bovine protein sequence database as the FASTA, including semi-tryptic peptides and specifying TOF measurement of fragment ions:

```
java -jar /usr/bin/MSGFPlus.20170113/MSGFPlus.jar -s mzMLs -mod Mods.txt  
-d 20180611-Ensemble-93-Bos_taurus.UMD3.1.pep.all.fasta -t 40ppm  
-m 0 -inst 2 -tda 1 -ntt 1
```

The Mods.txt file specified post-translational modifications to include in the search:

```
NumMods=2  
C2H2O2,C,opt,any,Carboxymethyl  
O1,M,opt,any,Oxidation
```

The search of all eight mzML files against the database (containing 22,118 sequences) completed in two hours on a quad-core Intel Core i7 processor. IDPicker 3.1⁵ then ingested the resulting mzIdentML files, applied a 2% FDR for the peptide-spectrum matches, and inferred a parsimonious set of proteins that explained at least two different peptide sequences. *All of these tools, I should note, are free to download and use.*

Most identification-based QC assessment starts from the principle “more is better.” But what value are we trying to maximize? One common answer is “sequence coverage.” Ideally, we could tile identified tandem mass spectra all the way across the protein sequence. On that note, begin with a look at the IDPicker user

interface:




















Accession	Count	Coverage	Distinct Peptides	Distinct Matches	Filtered Spectra
 2 (ENSBTAP00000022763.5)	1	 29	29	38	328
 6 (ENSBTAP00000016986.2)	1	 24	24	31	362
 11 (ENSBTAP00000009923.4)	1	 23	23	31	303
 3 (ENSBTAP00000019538.5)	1	 18	18	26	292
 10 (ENSBTAP00000023581.3)	1	 14	14	22	185
 1 (ENSBTAP00000010119.2)	1	 11	11	16	191
 5 (ENSBTAP00000006590.5)	1	 5	5	7	46
▶  4 (ENSBTAP00000004737.4,ENSBTAP00000033392.4)	2	18.29	3	4	33
 7 (ENSBTAP00000020243.2)	1	 3	3	4	19
▶  9 (ENSBTAP00000032384.3,ENSBTAP00000039139.3)	2	19.74	3	4	37
▶  8 (ENSBTAP00000015924.4,ENSBTAP00000052127.2)	2	12.80	2	2	11

Figure 3: IDPicker protein pane

How many proteins did this mixture contain? The vendor specifies six. Why then, does IDPicker show eleven distinguishable protein groups? Why do three of them look different (no Coverage display and two accession numbers)?

The first mystery should be easy to solve for anyone with proteomics experience. *No purified protein is ever 100% pure*. The first six rows represent the proteins specified by the manufacturer of this sample. The others represent “bonus” or “hitchhiker” proteins, depending on your perspective.

The other question can be addressed by the nature of protein sequence databases. Ensembl tries to include a protein sequence in the FASTA for each confident transcript known from RNA-Seq experiments. In some cases, that will mean that the same gene gives rise to multiple transcripts and thus different protein sequences that appear in the FASTA with different protein accessions. In other cases, that will mean that different paralogs (genes in a single organism that have duplicated from an ancestral gene) appear as highly similar sequences with different accession numbers. Tandem mass spectrometry measures peptides rather than proteins, so protein lists are inferred information. In cases like three of the lowest four proteins on the user interface, IDPicker could not discern which of the protein sequences is preferred to explain the identified peptides.

Sequence coverage is shown graphically in this interface, and yet the percentages are given when we export this view to a spreadsheet. From top to bottom, these values are as follows: 46%, 33%, 47%, 48%, 45%, 38%, 24%, 18%, 8%, 20%, 13%. I need to make an argument against the use of sequence coverage as a metric for quality control. First, protein sequences in FASTA files represent the full sequence translated at the ribosome. As a result, proteins that incorporate signal peptides will have their N-termini truncated; the protein lacks the first part of the FASTA protein sequence during most of its existence. Second, sequence coverage values tend to rise very slowly. Spectra identified to protein sequences are not distributed uniformly throughout the length of proteins. Some parts of the protein are “hot spots” for identification, and one is more likely to see a rare peptide (such as an enzymatic cleavage variant) that re-covers regions already seen than to observe a peptide for a never-before witnessed part of the protein. Third, if the protein always incorporates a post-translational modification or has a sequence that varies from the one in the sequence database, those regions will not be identified unless the search is tailored to take these factors into account.

For these reasons, IDPicker does not compute sequence coverage for each group or LC-MS/MS experiment provided by users.

The summary window is one of the more useful views for assessing identification success:

Group	Protein Groups	Distinct Peptides	Distinct Matches	Filtered Spectra
201203	10	85	108	308
201204	11	92	117	298
201205	11	81	98	232
201206	11	101	122	303
201207	11	75	92	236
201208	10	53	66	165
201209	11	61	75	152
201210	9	44	54	104

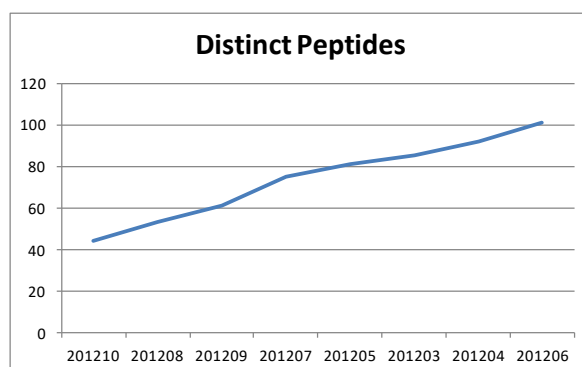
In this case, each LC-MS/MS experiment is given a shortened name representing when it was acquired. The protein groups column reveals that these eight experiments did not uniformly yield evidence for all proteins shown in the GUI; three experiments produced no spectra for at least one of the proteins. The protein group count is far more useful for QC in the context of highly complex protein mixtures, but for this “six protein” mixture, it is mostly useful in showing the “hitchhiker” proteins are not universally detected.

IDPicker uses a somewhat confusing vocabulary in its terms “distinct peptides, distinct matches, and filtered spectra.” The most conservative count is “distinct peptides,” which computes the numbers of distinct peptide sequences observed for a protein. I think of distinct peptides as the best proxy for biological information in a sample (unless post-translational modifications are the topic of the study). Note that spectra that represent different precursor charge state variants or different modifications of the same sequence will be lumped into one distinct peptide. This is different than the “distinct match” count. Charge state or modification differences accrue as different matches in the software. Finally we have the “filtered spectra” count, which is the total number of peptide-spectrum matches identified from an experiment. Since IDPicker works on the one spectrum-one peptide assumption, this measure is particularly useful when assessing the fraction of collected MS/MS scans that are identified by the bioinformatics framework.

We can rank each of these measures from least responsive to most responsive to variability. Protein group count is unlikely to move much from even a ten percent increase in sensitivity. The sum of protein coverage is also quite slow to respond. Distinct peptide sequence counts tend to move higher only when the amount of biological information in the measurement rises. Match counts and Spectral counts can jump under circumstances where the instrument is producing more identifiable spectra, but these additions may easily produce little gain in biological information. Imagine that someone has accidentally deactivated the “exclusion list” on an Orbitrap. As a result, the instrument collects far more identifiable spectra than is normal, and yet the number of distinct peptide sequences falls rather than rises.

With these thoughts in mind, let’s look at the table of the values for our eight LC-MS/MS experiments. How many distinct peptides did we find? The answer ranges from 44 in the final experiment to 101 in the June experiment. Is that more extreme than one should expect?

Alternately, would any of these experiments comprise an outlier on this basis? I would suggest plotting the sorted values. The result appears to the right. Is the line relatively straight, or does it curve down at the left and up at the right? The straightness suggests to me that the October and June experiments are quite consistent with the overall distribution. I do care about the ratio of best-to-worst, which in this case appears around 2.3. I don't like seeing these ratios rising above two, but the reality is that reproducibility in mass spectrometry commonly produces sensitivity differences of this magnitude. If this 2.3 ratio bothers you, try looking at the filtered spectra counts. In that field, this data set approaches a three-fold difference between maximum and minimum sensitivity. The other ratio that I frequently consider is the number of spectra identified divided by the number of distinct matches identified. The resulting ratio describes how much redundancy exists in the identified spectra, since spectra for the same match represent the same sequence, the same precursor charge state, and the same set of modifications. In these eight experiments, the value ranges from 1.93 in October to 2.85 in March.



Accelerating chromatogram examination through Skyline software

When it was introduced in 2010, Skyline¹¹ rapidly displaced a number of other software tools for targeted proteomics experiment design and / or analysis. Consequently, many proteomics laboratories assume it is only useful for Selected Reaction Monitoring (SRM) experiments or, in a pinch, analysis of SWATH experiments. Instead, this demonstration is intended to show that Skyline is quite useful for extracting chromatograms from “DDA” (Data Dependent Acquisition) or shotgun analyses of defined mixtures. The instructions will assume that Skyline has been installed but never used; if your copy gets a lot of exercise, you may have already changed some settings that will affect your results.

Begin by starting the Skyline software. Specify that you want to create a Blank Document. We will start by tuning the software in how it interprets sequences. First, Skyline anticipates the use of iodoacetamide, and our data reflect the use of iodoacetic acid. On the Settings menu, select “Peptide Settings.” Move over to the “Modifications” tab, and note that the software comes configured with a “Carbamidomethyl (C)” PTM. Uncheck this PTM in the list, and Edit the list to include one with these properties:

Name: Carboxymethyl (C)

Amino Acid: C

Chemical Formula: C₂H₂O₂

Variable: unchecked (We will assume all Cys are modified this way.)

The software may ask you about the UniMod PTM by the same name. It is fine to use the UniMod definition (these values were taken from that repository). Once you have okayed this addition to the list, you are returned to Peptide Settings. Now check “Carboxymethyl (C).” Click Okay to leave the Peptide Settings dialog.

Next, we will alter the Transition Settings, again accessible from the Settings Menu. On the “Filter” tab, we will specify that we are interested in producing precursor chromatograms. Replace the ‘y’ in

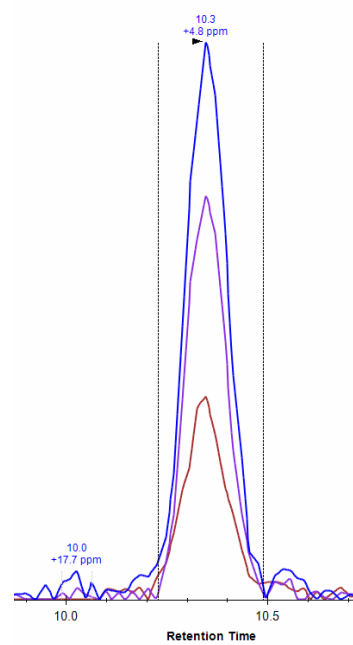
Ion Types with a 'p'. Move to the "Full-Scan" tab of the dialog box for the next step. Under "MS1 filtering," specify that we want to give the "Count" of Isotope peaks included. On Precursor mass analyzer, specify that precursors were measured in a "TOF." Click Okay to leave the Transition Settings dialog.

From the File menu, specify that you want to Import a FASTA file. As part of the course materials, you should have received a file named "20180817-six-bovine.fasta." Navigate to specify that file. You should find that the "Targets" list automatically fills with peptide sequences, sorted under the Ensembl accession for the six proteins in this mixture.

For our next step, we will see some magic! Return to the File: Import menu; this time we specify that we are ready to import some Results. Skyline accepts experiments in which each sample is represented by multiple LC-MS/MS experiments, but our data are the type already selected (single-injection replicates in files), so click Okay. Navigate to select all eight of the provided mzML files. The software will detect that their names share a common prefix: "2012." The names will be slightly shorter if you allow it to remove that prefix. When you allow the software to move ahead, get ready for a show; you will see it reading chromatograms from these mzMLs in real-time!

In reviewing our results, let's begin by clicking on the tab representing 201206. If you recall from the first section, the June replicate included the highest TIC measurements of any of the replicates. Click on the first protein name (the accession ending in "9923.4") and you should see chromatographic traces for all the peptides superimposed on each other. You can click and drag through a selected range of retention times to zoom to see them more clearly. You can get rid of the zoom by right-clicking on the chromatograms and selecting the last option ("Undo All Zoom / Pan").

We will revisit the set of peptides we examined manually in the first section of this document. The first protein contains the peptide beginning with DDG. Expand its information on the peptides list so you can confirm the m/z matches our table. When you click on the sequence, the plot of all peptide chromatograms for this protein is replaced by the information for this peptide alone. If you see a little arrow pointing to the top of one peak in the plot, it's an indicator that Skyline prefers this peak as representing this peptide. As we zoom closer, we see that three different traces rise and fall a little after 10 minutes in retention time. The legend shows that these ions all represent the doubly-charged precursor, with zero, one, or two extra neutrons. At the top of the peak, you can see the mass accuracy in ppm for this measurement. Is 4.8 ppm an acceptable variation in observed m/z for a QTOF instrument? Are any other possible retention times for this peptide a closer match in m/z?



I think we can agree that the DDG peptide is a well-behaved chromatogram. The peptide beginning with "IIAEGANG," again from the protein ending in "9923.4" will be a bit more difficult to interpret. Examine its chromatogram by clicking on this sequence. Do you accept the call Skyline has made for the peptide sequence, on the basis of mass error and coordinated rise of the three isotopes in MS

signal? We can see that the three traces rise and fall several times in the RPLC gradient. The reason is simple; many peptides are produced from this digestion, and unrelated sequences can produce intensity at an m/z that may appear to come from this peptide.

It seems clear that Skyline can greatly accelerate the process of checking particular peptides for mass accuracy and assessing chromatographic information. Given chromatography's prominence as a source of variance in proteomics, the chromatogram reconstruction automated by Skyline has particular value for QC. Skyline incorporates a variety of reports to help one evaluate the stability of chromatography among multiple replicates. Click on the DDG peptide again. This time use View: Retention Times: Replicate Comparison. The software will then visualize onscreen its best guess for the retention time of this peak in all of the LC-MS/MS experiments at once, showing the crest of each of the three traces in each of the eight files.

A similar view can be produced under View: Peak Areas: Replicate Comparison. We often think of the integration of intensity as a proxy for the quantity of a peptide, though this would only be comparable within the same peptide in a similar background. Instead of putting the peak areas side-by-side for each isotope, Skyline visualizes these areas in a single stacked bar graph. It uses a dot product ("idotp") to score how closely the relative areas of the peak traces correspond to the expected areas based on the "averagine" method¹² (if a library for this experiment is unavailable). A value close to 1.0 implies that the three isotopes have areas that conform with the expected ratios. An assessment of mass accuracy across files is similarly feasible in View: Mass Errors: Replicate Comparison.

In this case, we imported a FASTA file to enumerate automatically the set of peptide sequences that correspond to all six of the proteins. For routine use in a core facility, one could easily create a Skyline document specific to the peptides that are most useful to testing the instrument. Starting from this document, one can simply right-click on a peptide and select Delete to remove it from further consideration. With the right document ready to go, importing a new set of raw data to record retention time drift and mass accuracy in ppm for twenty peptides is a matter of a few minutes.

References

1. Paulovich AG, Billheimer D, Ham A-JL, Vega-Montoto L, Rudnick PA, Tabb DL, Wang P, Blackman RK, Bunk DM, Cardasis HL, Clauser KR, Kinsinger CR, Schilling B, Tegeler TJ, Variyath AM, Wang M, Whiteaker JR, Zimmerman LJ, Fenyo D, Carr SA, Fisher SJ, Gibson BW, Mesri M, Neubert TA, Regnier FE, Rodriguez H, Spiegelman C, Stein SE, Tempst P, Liebler DC. Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol Cell Proteomics*. 2010 Feb;9(2):242–254. PMID: PMC2830837
2. Bennett KL, Wang X, Bystrom CE, Chambers MC, Andacht TM, Dangott LJ, Elortza F, Leszyk J, Molina H, Moritz RL, Phinney BS, Thompson JW, Bunger MK, Tabb DL. The 2012/2013 ABRF Proteomic Research Group Study: Assessing Longitudinal Intralaboratory Variability in Routine Peptide Liquid Chromatography Tandem Mass Spectrometry Analyses. *Mol Cell Proteomics*. 2015 Dec;14(12):3299–3309. PMID: 26435129
3. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak M-Y, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, 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 EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol*. 2012 Oct;30(10):918–920. PMCID: PMC3471674
4. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*. 2003 Jul 1;31(13):3784–3788. PMCID: PMC168970
 5. Holman JD, Ma Z-Q, Tabb DL. Identifying proteomic LC-MS/MS data sets with BumberShoot and IDPicker. *Curr Protoc Bioinformatics*. 2012 Mar;Chapter 13:Unit13.17. PMCID: PMC4547833
 6. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. 2008 Dec;26(12):1367–1372. PMID: 19029910
 7. Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res*. 2007 Feb;6(2):654–661. PMCID: PMC2525619
 8. Lam H, Deutsch EW, Eddes JS, Eng JK, Stein SE, Aebersold R. Building consensus spectral libraries for peptide identification in proteomics. *Nat Methods*. 2008 Oct;5(10):873–875. PMCID: PMC2637392
 9. Frewen BE, Merrihew GE, Wu CC, Noble WS, MacCoss MJ. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal Chem*. 2006 Aug 15;78(16):5678–5684. PMID: 16906711
 10. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun*. 2014;5:5277. PMID: 25358478
 11. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*. 2010 Apr 1;26(7):966–968. PMCID: PMC2844992
 12. Senko MW, Beu SC, McLaffertycor FW. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J Am Soc Mass Spectrom*. 1995 Apr;6(4):229–233. PMID: 24214167