Published in final edited form as:

J Proteome Res. 2016 August 05; 15(8): 2500–2509. doi:10.1021/acs.jproteome.6b00021.

BatMass: a Java software platform for LC/MS data visualization in proteomics and metabolomics

Dmitry Avtonomov¹, Alexander Raskind², and Alexey I. Nesvizhskii^{*,1,3}

¹Department of Pathology, University of Michigan, Ann Arbor, MI 48109

²BRCF Metabolomics Core, University of Michigan, Ann Arbor, MI 48109

³Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109

Abstract

Mass spectrometry (MS) coupled to liquid chromatography (LC) is a commonly used technique in metabolomic and proteomic research. As the size and complexity of LC/MS based experiments grow, it becomes increasingly more difficult to perform quality control of both raw data and processing results. In a practical setting, quality control steps for raw LC/MS data are often overlooked and assessment of an experiment's success is based on some derived metrics such as “the number of identified compounds”. Human brain interprets visual data much better than plain text, hence the saying “a picture is worth a thousand words”. Here we present BatMass software package which allows to perform quick quality control of raw LC/MS data through its fast visualization capabilities. It also serves as a testbed for developers of LC/MS data processing algorithms by providing a data access library for open mass spectrometry file formats and a means of visually mapping processing results back to the original data. We illustrate the utility of BatMass with several use cases of quality control and data exploration.

Keywords

Mass spectrometry; LC/MS; Data Visualization; Java

Introduction

Liquid chromatography mass spectrometry (LC/MS) has long become a routine investigation method in various fields of bioanalysis, such as proteomics and metabolomics. Any study utilizing LC/MS begins with processing of raw mass-chromatograms from the instruments. Data needs to be checked for quality – performance of the LC system, stability of m/z traces over time, carryover contamination from previous runs¹. Then useful information needs to be extracted in the form of m/z values and their corresponding intensities over time (LC/MS features)². These are critical steps which are often not given

Corresponding author: Alexey I. Nesvizhskii, Department of Pathology, University of Michigan, 4237 Medical Science I, Ann Arbor, MI, 48109, nesvi@med.umich.edu, Tel: +1 734 764 3516.

enough attention as they are tedious - there is no easy way to check the stability of measured masses or validate that feature finding algorithm did a good job detecting LC/MS features.

As capabilities of LC/MS instruments improve and control software allows for better automation of data acquisition process, less experienced users gain wider hands-on access to this technology. They often treat mass spectrometers as tools that “just work”, which is not always the case. It takes time and experience to learn how to assess the quality of an LC/MS run by looking at total ion chromatograms and individual spectra – most common visualizations of LC/MS data, provided by instrument vendors as well as open source software. For some reason, LC/MS data visualization is not on par with instrumentation. Not often do investigators take a closer look at their own raw data, with the most commonly used metric for the quality of a dataset in proteomics, for example, being the number of identifications at peptide or protein level, and the number of detected and identified features in metabolomics³⁻⁵. Even though specialized software packages exist designed specifically for quality control (QC) of LC/MS data⁶⁻¹⁰, their output is condensed into several QC metrics values and dataset-wide distribution plots of those metrics. While they provide useful information, these tools are seldom used in the proteomics/metabolomics community, at least not at an early stage of the data analysis, and the standard way of assessing data quality is still through examination of ion chromatograms and identification rates. Comprehensive 2D visualization of raw LC/MS data (in m/z and RT dimensions), on the other hand, can provide a detailed insight into the quality of chromatographic separation, average peak elution times, stability of measured masses over time and quality of LC/MS feature detection even to an unexperienced user with a minimal training.

Recent emergence of novel data-independent acquisition techniques (DIA) such as SWATH¹¹, MSe¹², pSMART¹³, WiSIM¹³, presents its challenges to investigators trying to come up with optimal acquisition strategies and processing algorithms. In DIA, unlike conventional data-dependent acquisition (DDA), precursors are not isolated for fragmentation selectively, but instead are co-fragmented using wide isolation windows. A more complete list of DIA methods can be found in recent reviews on the topic^{14,15}. DIA method optimization would be simplified if there was a way to visualize the data properly, however, despite the growing popularity of DIA, development of appropriate visualization tools is lagging behind. The same is true for LC/MS feature detection algorithms development and applications – the most important step in quantitative “-omics” experiments. Inspection of feature detection results should be an easy task, considering that tweaking of multiple parameters might be required for each particular dataset, affecting the quality and confidence of quantification, but commonly the only visualization available to the user is an extracted ion chromatogram, which is a one-dimensional representation of the complex three-dimensional LC/MS feature (m/z, retention time, intensity). Viewing the data in three dimensions, on the other hand, clearly reveals which features were detected correctly and, more importantly - which ones were missed.

There are existing open source software packages which provide helpful viewers for MS data, such as TOPPView from OpenMS¹⁶, MZMine2¹⁷, Mass++¹⁸, but being all specialized in particular aspects of MS data processing, they are not as useful for visual QC purposes. The software package BatMass, presented in this paper, was designed to fill that niche and

provide users with capability to quickly explore raw LC/MS data and get easy visual links from processed results back to it. It is also the only visualization tool capable of displaying LC/MS DIA data in 2D.

Methods

Implementation

BatMass is an open-source Java program written using the NetBeans Platform (<http://platform.netbeans.org>). The NetBeans Platform is a modular framework, which means any application can be extended by providing additional plugins (modules) that can be added to existing application without the need to reinstall the software. It also provides the base for creating a feature-rich graphical user interface on par with commercial desktop applications. Existing data processing algorithms or visualizations can be hooked up to the system as plugins using provided extension points. Writing the plugins does not require thorough knowledge of the whole existing code base of the application, but rather understanding of NetBeans Platform and documentation for extension points of existing application modules. Some parts of compomics-utilities package¹⁹ were used in the software. There are no specific hardware requirements, however larger amounts of RAM are desirable. Depending on the distribution of scans in mzML and mzXML files between MS1 and MS2, data compression applied and floating point precision used as much memory as the original size of the file might be required if all the scans are MS1 only. Most of the time though, it should be possible to run with available RAM just being a fraction of the file size.

MS data access layer

As BatMass was written in Java, it needed a way to access LC/MS data files in open formats (mzML and mzXML) using native Java libraries, however, to our knowledge, there is only one such library – JmzReader²⁰. Open XML based mass-spectrometry data formats cannot always be reliably read using standard-conforming readers (such as JmzReader), because real-life data files, e.g. converted with older versions of vendor specific software from native vendor formats, not always follow the standards and thus can't be read. The situation in the mass-spectrometry world is reminiscent of that in the world of internet, when some browsers are not able to render a webpage correctly because it uses some non-standard markup. However, in the world of the internet the web pages normally try to use the syntax understood by the current browsers, while in the world of mass-spectrometry it's the other way round - software that has to access MS data tries to guess how to properly read it. We have found JmzReader to not be fault tolerant enough, for example it was not able to read Thermo RAW files converted to mzXML using the popular ReAdW program (Thermo RAW to mzXML converter), which is still being used, and other ill formed mzML/mzXML files. Furthermore, we found it too slow for interactive data exploration and visualization. Vendor specific file formats can be converted to mzML/mzXML using ProteoWizard²¹ (more details available²²). A custom data access library was developed for BatMass to fulfill the speed requirements (parsing speed is comparable to the C++ implementation from OpenMS) and to automate memory management. It provides a rich API for accessing scan meta-data and spectra, including support for MS-Numpress compression²³ in mzML files. As the API is

separated from the implementation, it is possible to add support for other file formats as well.

The library is capable of accessing conventional Data Dependent Acquisition (DDA) data as well as newer Data Independent Acquisition (DIA) runs. Data formats for representing DIA experiments are not well established yet, and oftentimes the full information about precursor isolation windows in DIA is not recorded in the files. In such cases using simple heuristics the library tries to guess if the data was acquired using a DIA strategy and groups MS2 scans coming from the same precursor windows.

Data organization and visualization

Out of the box BatMass provides a project system for maintaining files in logical groups (Fig. 1). Like everything else in BatMass, the project system is extendable; developers can create new project types, which are meant to provide different sets of actions, applicable in different contexts. For example, in a proteomics setting an action for searching tandem mass spectra (MS/MS spectra) against a protein database might be available for raw files, while such an action in a metabolomics project might provide options for searching against a spectral library.

With the goal of visualizing raw LC/MS data and overlaying processing results on top, BatMass provides a number of data viewers. New ones can be added by developers using extension points. No LC/MS viewing tool can go without spectrum and chromatogram viewers, which are provided, but the main point of interest is the 2D map view. 2D viewer is a two dimensional heat map of m/z vs. RT with color coding for signal intensity. It is the most powerful tool, which is not always found in similar software packages. The tools mentioned above (OpenMS, Mass++, MZMine2) offer this type of data view, but they either do not provide the same functionality or are much slower, rendering them less useful for exploratory data analysis. MS vendor software normally does not provide that type of view either, or it is very limited. To our knowledge, BatMass is also the only package capable of visualizing fragmentation spectra from DIA experiments in 2D, allowing the user to view it in the same way as regular MS1 scan maps. This feature is very helpful in assessing and optimizing specific flavors of DIA acquisition, it is also the only one providing an insight into the quality of measured DIA MS2 spectra with regards to m/z scan to scan stability (Fig. 2).

The main extension points in the system are two tabular viewers, which display data as either a simple table or as a tree-table, which is a table with the left most column representing a tree, while other columns display plain tabular data for the corresponding row. These tools can be used for any sort of tabular data that might be linked to raw LC/MS files, for example: identified peptides in the form of .pep.xml files in proteomics, detected LC/MS features (presented as retention time spans and m/z values) in custom file formats in proteomics and metabolomics, etc. By providing parsers for custom file formats and converters to viewers' internal data models new data mapping capabilities can be introduced. For example, it is possible to create a parser for a custom spectral library file format and it will automatically be possible to overlay the whole library onto a 2D view.

One important feature of BatMass is viewer synchronization; any viewer can be linked to any other viewer using drag and drop functionality. This allows, for example, to view the same regions of different runs in a 2D map view, while zoom and pan events will be synchronized between the viewers. It also allows seamlessly navigate from identifications (peptides, metabolites, etc.) in a tabular view directly to the corresponding spectrum at the LC apex or a 2D map view of the whole LC/MS feature with a double click on the corresponding table row.

The layout of windows within the application is completely flexible. They can be docked in various positions, minimized, made to slide from the sides or undocked and float separately from the main window (Fig. 1).

Results

2D map visualization of an LC/MS run is the most informative compared to other modes of visualization. It has the capability to correctly render profile and centroided data without any user-selectable parameters. We have chosen not to provide a 3D viewer in BatMass. 3D scenes are surely nice looking, but are very hard to navigate with the mouse, which is inherently a 2D tool. That is the reason why BatMass does not have support for 3D and instead focuses on providing a comprehensive 2D visualization experience. Unlike viewing single spectra, viewing the whole LC/MS run in 2D is a memory/CPU-intensive task, as it requires access to all spectra at a given MS level at once.

Comparison of 2D visualizations

To compare BatMass to other tools capable of generating 2D views, we've used the publicly available synthetic phospho-peptide dataset²⁴ (available at <http://proteomexchange.org> under ID PXD000138). On a typical desktop computer (Intel Core-i5 2400 quad-core CPU, 8GB RAM, 7200 RPM HDD), a single raw file (5.RAW from the above dataset) was converted to mzXML format (64-bit for both m/z and intensity values, no compression) which resulted in a 1.7Gb output file.

MZMine2 took over 2 minutes to load the file and provided no simple way to navigate around the run. It comes with few predefined color schemes, which are not well suited to viewing high dynamic range data. Tweaking the color scheme took another 30 seconds and had to be done every time a new file was opened. Zooming took considerable time (tens of seconds) and zoomed-in views did not show fine details, as data were binned to 0.01 m/z. TOPPView from OpenMS performed initial data import faster (~1min), and provided real-time navigation around the run with the mouse. However, it does not interpolate values over retention time axis and m/z axis is binned to the same 0.01 m/z bins as in MZMine2, leaving out the fine details. In Mass++ the initial data structure was imported in just 20sec, but reading the spectra and building 2D map took over 5min. As each zoom event requires re-parsing of the data from the file, it makes locating the desired LC/MS region a long and tiresome process. Unlike MZMine2 and TOPPView though, the data are not binned, but instead somehow interpolated, and even for centroided spectra only the interpolated image can be viewed. Mass++ also could not read some mzXML files, which other tools could.

In contrast, BatMass took 30sec from click to a 2D view being displayed, navigation is real-time and can be done using either a mouse or a Go-To dialog akin to Go-To dialogs in text processors, where the user can specify the exact m/z -RT region to be displayed. Automatic dynamic range scaling enables viewing data of different intensity ranges without changing any parameters, and viewer settings are persisted between sessions. As the viewer does not use binning, it is possible to zoom in arbitrarily close in m/z dimension, revealing fine m/z variations.

Manual QC of raw data

Assessing the quality of raw LC/MS data is a multistep process. As a part of standard routine, total ion chromatograms (TICs) or base peak chromatograms (BPCs) can be examined to reveal any “abnormalities”, i.e. poor peak separation in LC, abnormal peak shapes, abrupt intensity drops (e.g. when a bubble forms on the electrospray (ESI) needle), column overload, etc. Spectra might be checked as well for presence of isotopic clusters. However, it's hard to assess the quality from looking at individual spectra as possible variations in accuracy of mass measurement are not apparent. 2D map view provides a quick and easy way to verify mass stability (how stable m/z traces remain over time) as an additional important quality parameter (Fig. 3). When m/z values are not stable over time, it is much harder for feature finding algorithms to properly detect LC/MS features and correctly determine the masses, which prevents proper compound identifications and leads to incorrect quantitation. A 2D view is also very helpful in detection of glitches or strange results during LC/MS data preprocessing, e.g. centroiding (Fig. 3).

DIA method development

Data independent acquisition starts gaining momentum, both in proteomics¹⁴ and metabolomics²⁵, as more instruments support that mode of acquisition. Different DIA data analysis strategies are being tried with varying degree of success⁸. In our own work, we have found it very helpful to visualize DIA MS2 data using 2D maps to troubleshoot problems with pilot experiments (Figs. 3, 4). Visualization of fragmented m/z swaths for the whole run allows, for example, to quickly check how many precursor ions fall into a single swath on average, examine fragmentation rate to check how many precursors were left unfragmented, to check if the cycle time was adequate for the chromatography protocol used, how good MS2 results were in general, and if there is a strong correlation between precursors and fragments – the feature particularly important for untargeted computational tools for DIA data such as DIA-Umpire²⁶ (Fig. 2).

Manual QC of downstream processing results

The main pieces of information extracted from modern LC/MS experiments are LC/MS features - elution profiles of ions of particular m/z values. This is especially true for experiments that do not use fragmentation information, as reliable masses and retention times (RTs) are the only variables that identify compounds in this case. This is a commonplace in metabolomics, where fragmentation of small molecules often yields only a few peaks and thus less informative^{27,28}. In such a scenario – identification of metabolites that is largely driven by m/z and RT based on a pre-built library of known compounds – it becomes particularly important to evaluate, and ideally optimize, the performance of the

feature detection algorithm applied to the data. This means not only checking the quality of detected features, but also figuring out if any significant LC/MS features have been missed. The check for what has been missed is especially important but, in general, hard to implement in practice. Unlike a detected feature, the quality of which can be assessed, for example, by plotting its XIC, there is no simple test for what has not been found in the first place. This requires a ground truth dataset, where all the features are known beforehand, so such a dataset must be either generated computationally or a human expert needs to label all the signal-containing regions of an LC/MS run manually. With BatMass it is possible to overlay feature detection results over 2D map view to get the visual confirmation of an algorithm's success rate.

For demonstration purposes, we have run two different feature detection algorithms (XCMS²⁹ and Agilent MassHunter) configured with different parameters (intensity cutoffs, numbers of isotopes for a feature to be accepted, etc.) against a single untargeted metabolomics LC/MS run of pooled human plasma sample analyzed on an Agilent 6530 Q-TOF instrument. The features detected by these two software algorithms were overlaid over a 2D map of the run, see Fig. 5. The plot shows that the differences one might get by using different feature detection software tools and different algorithm parameters can be very significant. BatMass helps to compare the algorithms' results to each and to visualize the differences. This, in turn, helps to identify the effects of various parameter settings, and thus, alone or in combination with computational solutions^{30,31}, can aid the parameter selection process for a particular feature detection algorithm or can be used in conjunction with software that assesses data quality computationally, e.g. msCompare³².

Via the plugin system BatMass allows overlay of custom data over the 2D map, which can be LC/MS features from a specific file format (e.g. XCMS or Agilent MassHunter file formats are currently supported for metabolomics), identifications (e.g. pepXML files for proteomics), or anything else that can be represented in m/z - retention time coordinates. As an example application of BatMass, during the development of a software tool DIA-Umpire for DIA MS data we needed a tool to visualize critical steps of raw data analysis: feature finding, isotopic grouping and grouping of precursor-fragment features. An integration layer for DIA-Umpire has been quickly implemented and used to optimize the DIA-Umpire algorithm. BatMass turned out to be indispensable for identifying cases of signals missed by the DIA-Umpire's feature detection module, as well as signals that were split into multiple ones in retention time, to figure out what peculiarities of the signal have thrown the algorithm away. In another example, we are using BatMass as part of the ongoing work to improve the metabolomics data analysis workflows, e.g. by using BatMass to visualize and interpret initially unidentified features found to have, according to the Correlation Calculator (a recent addition to the MetScape³³ suite of computational metabolomics tools; <http://metscape.ncibi.org/calculator.html>), a high correlation of their quantitative profiles across multiple samples to some of the identified features. We believe that at present BatMass remains the only tool that allows quick development of plug-in parsers for custom LC/MS feature storage formats and the import and overlay of the data over a 2D view - the functionality that is needed for advanced MS computational tool development such as examples mentioned above.

Library-based targeted experiments

Targeted MS-based studies³⁴ are the area where BatMass can offer great help to the researchers. A common requirement is extraction of signals from raw data based on a pre-built library of known compounds, containing corresponding annotations, masses and retention times. Software packages exist for library based XIC extraction, e.g. SkyLine³⁵. However, when signals are weak or completely absent, it is often hard to figure out if a signal was present or not. The standard approach is to extract all the signals in the region of interest, even if it's only noise; it's difficult to verify such events using only chromatogram and spectrum viewers, while the answer might be more easily discernible when viewed in 2D.

LC/MS data are relatively noisy and, depending on the width of the extraction window being used, noise might be integrated and counted as meaningful signal. This often happens in targeted metabolomics and proteomics when several LC/MS runs are being compared. The lists of extracted LC/MS features from different runs rarely match perfectly – some features are often detected only in a subset of runs and are marked as missing in the rest of the samples. It is common in such situation to try and “fill the gaps” by revisiting runs for which the value was missing and blindly integrating the signal from a particular range of masses and retention times^{1,36}. However, one should be careful with such an approach because the signal might simply not be there at all (Fig. 6). Targeted visualization of selected features of interest is important in untargeted metabolomics and proteomics studies as well. Despite many advanced LC/MS alignment algorithms that have been described in the literature³⁶, computational tools alone would never be able to completely address this problem. Thus, in those cases where additional level of confidence in the accuracy of detection and quantification of a particular feature of interest is desired (e.g. a feature determined to be a candidate biomarker based on the downstream analysis of the entire dataset), BatMass can assist with manual confirmation of the results.

Another common example is validation of presence of a compound in the sample by extracting chromatograms for a particular m/z value and comparing positions of elution peaks to a pre-built library of compound masses and retention times. Depending on the width of m/z extraction window, multiple chromatographic peaks might often be detected in a single such XIC. In this case viewing data in 2D is much simpler and more informative than looking through spectra, providing a clear bird's eye view of the situation. Masses in each elution peak might be slightly different suggesting that those peaks relate to different chemical compounds or the same compound but of slightly different structure. In case of peptides such a situation might arise when the same post translational modification is attached to different sites in the backbone, leading to a difference in chromatographic retention, but not the mass (Fig. 7).

Conclusions

Visualization tools are indispensable for analysis of any data. The software package BatMass described in this work provides a set of visualizations for traditional mass-spectrometry data as well as emerging DIA acquisition strategies. It provides standard spectrum and chromatogram (TIC, base peak, extracted ion chromatograms) viewers, but the

most powerful is the 2D map viewer. Unlike commonly implemented in other software, it does not bin the data, which is required to quickly assess quality of LC/MS runs and scan-to-scan mass stability. It has automated dynamic range scaling, allowing one to clearly see even the weakest, near noise level signals. LC/MS feature finding results, as well as peptide or metabolite identification results can be overlaid on top of spectrum and 2D map views. With the 2D map tool it is also easy to check quantitation results. Instead of using extracted ion chromatograms, which is error prone and tedious as single spectra need to be checked manually for co-eluting ion species, mapping identifications back to raw data in 2D gives a clear answer at a glance. Viewers can be linked together to allow quick navigation between feature-finding or identification results and raw data (spectra and LC/MS regions in 2D). BatMass is also very useful for development of feature-finding and targeted identification algorithms, as results can be overlaid on top of raw data visual map, where it is easy to assess the performance of the processing algorithm.

We will be adding new features to the software over time, including overlay of MS/MS events and identifications, marking of isolation windows for DIA data, better denoising capabilities and contoured feature plotting instead of bounding boxes. The data access library will be expanded with signal processing algorithms for denoising and feature extraction. The updates do not require reinstallation of BatMass, the already installed instance is updated instead. Upon every start of the application it automatically checks for updates, which are delivered through GitHub. The update policy can be changed in the settings.

The software and user manual as well as developer starter-guides are available at the website <http://batmass.org>. The source code is available under Apache 2.0 license and is hosted on GitHub at <https://github.com/chhh/batmass> for the core BatMass code and <https://github.com/chhh/msftbx> for the data-access library.

Acknowledgments

We would like to acknowledge Charles Burant for useful discussions and providing access to metabolomics mass spectrometry files, Venky Basrur for providing sample proteomics mass spectrometry data, Alla Karnovsky, Chih-Chiang Tsou and Sub Pennathur for useful discussions. This work was supported in part by grant U24 DK097153 of NIH Common Funds Project to the University of Michigan (the Michigan Comprehensive Metabolomics Research Core, MRC2) and by NIH Grant R01-GM-094231 (to A.I.N).

References

1. Bereman MS. Tools for monitoring system suitability in LC MS/MS centric proteomic experiments. *Proteomics*. 2015; 15:891–902. [PubMed: 25327420]
2. America AH, Cordewener JH. Comparative LC-MS: a landscape of peaks and valleys. *Proteomics*. 2008; 8:731–749. [PubMed: 18297651]
3. Wang X, Chambers MC, Vega-Montoto LJ, Bunk DM, Stein SE, Tabb DL. QC Metrics from CPTAC Raw LC-MS/MS Data Interpreted through Multivariate Statistics. *Anal Chem*. 2014; 86:2497–2509. [PubMed: 24494671]
4. Rudnick PA, Clauser KR, Kilpatrick LE, Tchekhovskoi DV, Neta P, Blonder N, Billheimer DD, Blackman RK, Bunk DM, Cardasis HL, Ham AJ, Jaffe JD, Kinsinger CR, Mesri M, Neubert TA, Schilling B, Tabb DL, Tegeler TJ, Vega-Montoto L, Variyath AM, Wang M, Wang P, Whiteaker JR, Zimmerman LJ, Carr SA, Fisher SJ, Gibson BW, Paulovich AG, Regnier FE, Rodriguez H, Spiegelman C, Tempst P, Liebler DC, Stein SE. Performance metrics for liquid chromatography-

tandem mass spectrometry systems in proteomics analyses. *Mol Cell Proteomics*. 2010; 9:225–241. [PubMed: 19837981]

5. Tabb DL. Quality assessment for clinical proteomics. *Clin Biochem*. 2013; 46:411–420. [PubMed: 23246537]
6. Ma ZQ, Polzin KO, Dasari S, Chambers MC, Schilling B, Gibson BW, Tran BQ, Vega-Montoto L, Liebler DC, Tabb DL. QuaMeter: multivendor performance metrics for LC-MS/MS proteomics instrumentation. *Anal Chem*. 2012; 84:5845–5850. [PubMed: 22697456]
7. Taylor RM, Dance J, Taylor RJ, Prince JT. Metriculator: quality assessment for mass spectrometry-based proteomics. *Bioinformatics*. 2013; 29:2948–2949. [PubMed: 24002108]
8. Walzer M, Pernas LE, Nasso S, Bittremieux W, Nahnsen S, Kelchtermans P, Pichler P, van den Toorn HW, Staes A, Vandenbussche J, Mazanek M, Taus T, Scheltema RA, Kelstrup CD, Gatto L, van Breukelen B, Aiche S, Valkenburg D, Laukens K, Lilley KS, Olsen JV, Heck AJ, Mechtler K, Aebersold R, Gevaert K, Vizcaino JA, Hermjakob H, Kohlbacher O, Martens L. qcML: an exchange format for quality control metrics from mass spectrometry experiments. *Mol Cell Proteomics*. 2014; 13:1905–1913. [PubMed: 24760958]
9. Simader AM, Kluger B, Neumann NKN, Bueschl C, Lemmens M, Lirk G, Kraska R, Schuhmacher R. QCScreen: a software tool for data quality control in LC-HRMS based metabolomics. *BMC Bioinf*. 2015; 16:1–9.
10. Perez-Riverol Y, Xu QW, Wang R, Uszkoreit J, Griss J, Sanchez A, Reisinger F, Csordas A, Ternent T, del-Toro N, Dianes JA, Eisenacher M, Hermjakob H, Vizcaino JA. PRIDE Inspector Toolsuite: moving towards a universal visualization tool for proteomics data standard formats and quality assessment of ProteomeXchange datasets. *Mol Cell Proteomics*. 2015; 5:305–317.
11. Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, Reiter L, Bonner R, Aebersold R. 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*. 2012; 11:O111.016717.
12. Silva JC, Denny R, Dorschel CA, Gorenstein M, Kass IJ, Li GZ, McKenna T, Nold MJ, Richardson K, Young P, Geromanos S. Quantitative proteomic analysis by accurate mass retention time pairs. *Anal Chem*. 2005; 77:2187–2200. [PubMed: 15801753]
13. Prakash A, Peterman S, Ahmad S, Sarracino D, Frewen B, Vogelsang M, Byram G, Krastins B, Vadali G, Lopez M. Hybrid data acquisition and processing strategies with increased throughput and selectivity: pSMART analysis for global qualitative and quantitative analysis. *J Proteome Res*. 2014; 13:5415–5430. [PubMed: 25244318]
14. Sajic T, Liu Y, Aebersold R. Using data-independent, high-resolution mass spectrometry in protein biomarker research: perspectives and clinical applications. *Proteomics: Clin Appl*. 2015; 9:307–321. [PubMed: 25504613]
15. Chapman JD, Goodlett DR, Masselon CD. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrom Rev*. 2014; 33:452–470. [PubMed: 24281846]
16. Sturm M, Kohlbacher O. TOPPView: an open-source viewer for mass spectrometry data. *J Proteome Res*. 2009; 8:3760–3763. [PubMed: 19425593]
17. Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinf*. 2010; 11:395.
18. Tanaka S, Fujita Y, Parry HE, Yoshizawa AC, Morimoto K, Murase M, Yamada Y, Yao J, Utsunomiya SI, Kajihara S, Fukuda M, Ikawa M, Tabata T, Takahashi K, Aoshima K, Nihei Y, Nishioka T, Oda Y, Tanaka K. Mass++: A Visualization and Analysis Tool for Mass Spectrometry. *J Proteome Res*. 2014; 13:3846–3853.
19. Barsnes H, Vaudel M, Colaert N, Helsens K, Sickmann A, Berven FS, Martens L. compomics-utilities: an open-source Java library for computational proteomics. *BMC Bioinf*. 2011; 12
20. Griss J, Reisinger F, Hermjakob H, Vizcaino JA. jmxReader: A Java parser library to process and visualize multiple text and XML-based mass spectrometry data formats. *Proteomics*. 2012; 12:795–798. [PubMed: 22539430]
21. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*. 2008; 24:2534–2536. [PubMed: 18606607]

22. Holman JD, Tabb DL, Mallick P. Employing ProteoWizard to Convert Raw Mass Spectrometry Data. *Curr Protoc Bioinformatics*. 2014; 4613(24):11–19.
23. Teleman J, Dowsey AW, Gonzalez-Galarza FF, Perkins S, Pratt B, Rost HL, Malmstrom L, Malmstrom J, Jones AR, Deutsch EW, Levander F. Numerical compression schemes for proteomics mass spectrometry data. *Mol Cell Proteomics*. 2014; 13:1537–1542. [PubMed: 24677029]
24. Marx H, Lemeer S, Schliep JE, Matheron L, Mohammed S, Cox J, Mann M, Heck AJ, Kuster B. A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat Biotechnol*. 2013; 31:557–564. [PubMed: 23685481]
25. Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, Kanazawa M, VanderGheynst J, Fiehn O, Arita M. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods*. 2015; 12:523–526. [PubMed: 25938372]
26. Tsou CC, Avtonomov D, Larsen B, Tucholska M, Choi H, Gingras AC, Nesvizhskii AI. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods*. 2015; 12:258–264. [PubMed: 25599550]
27. Johnson CH, Ivanisevic J, Benton HP, Siuzdak G. Bioinformatics: The Next Frontier of Metabolomics. *Anal Chem*. 2015; 87:147–156. [PubMed: 25389922]
28. Cho K, Mahieu NG, Johnson SL, Patti GJ. After the feature presentation: technologies bridging untargeted metabolomics and biology. *Curr Opin Biotechnol*. 2014; 28:143–148. [PubMed: 24816495]
29. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal Chem*. 2006; 78:779–787. [PubMed: 16448051]
30. Uppal K, Soltow QA, Strobel FH, Pittard WS, Gernert KM, Yu T, Jones DP. xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. *BMC Bioinf*. 2013; 14:1–12.
31. Libiseller G, Dvorzak M, Kleb U, Gander E, Eisenberg T, Madeo F, Neumann S, Trausinger G, Sinner F, Pieber T, Magnes C. IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinf*. 2015; 16:1–10.
32. Hoekman B, Breitling R, Suits F, Bischoff R, Horvatovich P. msCompare: A Framework for Quantitative Analysis of Label-free LC-MS Data for Comparative Candidate Biomarker Studies. *Mol Cell Proteomics*. 2012; 11
33. Karnovsky A, Weymouth T, Hull T, Tarcea VG, Scardoni G, Laudanna C, Sartor MA, Stringer KA, Jagadish HV, Burant C, Athey B, Omenn GS. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics*. 2012; 28:373–380. [PubMed: 22135418]
34. Ebhardt HA, Root A, Sander C, Aebersold R. Applications of targeted proteomics in systems biology and translational medicine. *Proteomics*. 2015; 15:3193–3208. [PubMed: 26097198]
35. 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; 26:966–968. [PubMed: 20147306]
36. Sandin M, Teleman J, Malmström J, Levander F. Data processing methods and quality control strategies for label-free LC–MS protein quantification. *Biochim Biophys Acta*. 2014; 1844:29–41. [PubMed: 23567904]

Abbreviations

LC/MS	liquid chromatography mass spectrometry
TIC	total ion chromatogram
XIC	extracted ion chromatogram
DIA	data independent acquisition

DDA	data dependent acquisition
ESI	electro spray ionization
SWATH	sequential window acquisition of all theoretical mass spectra
QC	quality control
2D	two-dimensional
RT	retention time

Figure 1. Main application window hosting the project explorer and different data viewers
Files are organized in projects, logical subdirectories thereof, and can be set attached as child nodes to other files. Actions are provided via context menus. All windows within the main frame can be moved freely around to create a workspace convenient to the user.

Figure 2. Two synchronized 2D map viewers (MS1 and MS2 scans) displaying the same LC/MS DIA data file

X axis: m/z ; Y axis: retention time in minutes. The left panel is set up to display MS1 spectra and the viewport is limited to show only the m/z range of a single SWATH precursor isolation window (649-675 m/z in this example). The right panel is set to display MS2 fragmentation spectra from that MS1 window only, with the whole fragment m/z range shown. As the panels are displaying data at different MS levels (MS1 vs. MS2), they are only synchronized with respect to retention time axis. Note how intense precursors on the left align with series of fragments on the right – an important characteristic of DIA data. Data: UPS1 proteins mixed with Human cell lysate, AB Sciex TripleTOF 5600.

Figure 3. Real life examples of problematic raw MS data easily detectable using 2D visualization (Top row) LC/MS run performed on a Thermo Scientific Orbitrap Fusion instrument using suboptimal implementation of a novel WiSIM DIA strategy: MS1 acquisition is segmented and MS2 scans are acquired with 12Da windows. Left to right - consecutive zoom-in views to region 883-886m/z, 42-48 min. From the overview image on the left everything seems to be fine, lots of fragments are visible and almost no unfragmented precursors are left in the isolation window m/z region. But when zoomed in closer on a single isotopic cluster, the problem is revealed. Each m/z trace is highly unstable and scan-to-scan jumps of over 100ppm are observed. A possible explanation is that SIM scans (Thermo instrument specific setting for the particular type of scan), that were used to acquire fragmentation spectra, might not be adequate for this application. (Bottom left) A different example if m/z trace instability over time. The m/z trace first appears at 281.2400, then as intensity goes up it

shifts up to 281.2700 (105ppm difference), and then stabilizes at m/z 281.2485 (75ppm from the m/z at elution apex). Data: Human plasma, Agilent 6530. **(Bottom right)** Unusual centroiding behavior. Centroided (left) and profile (right) data from the same LC/MS run performed on a Thermo Orbitrap Fusion are shown. Centroiding was done with ProteoWizard, using “prefer vendor peak picking” option. Profile data looks like the m/z trace is very stable, however it was split into two parallel traces 10ppm apart after centroiding.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 4. 2D map of MS2 spectra from the same precursor window in a SWATH-MS DIA experiment

Standard SWATH 25Da isolation windows were used, the spectra come from precursor window 700-725 m/z, the window itself is readily seen in the picture as a swath of m/z containing unfragmented precursors. Many signals (yellow to red) are observed in the first half of the run inside the window, which means a lot of precursor ions were not fragmented. Data: Human cell lysate, ABSciex TripleTOF 5600.

Figure 5. 2D comparison display of LC/MS features detected by Agilent MassHunter (Molecular Feature Extractor) and XCMS (Massifquant) with different parameters
 MassHunter was only allowed to pick up a signal if at least two isotopic peaks could be detected in a feature, while XCMS was detecting single mass traces without any additional filtering applied. Data: Human plasma, Agilent 6530.

Figure 6. 2D visualization of selected features across multiple runs

(**Top row**) Synchronized 2D views of three LC/MS runs. Bottom row: corresponding extracted ion chromatograms (XICs) for the monoisotopic peak, extraction window width 30ppm. Only the LC/MS run in the right panel contains non-noise signal, however, it is unclear from the XICs alone. Even though no LC peak is visible in the XICs, the total area under the curve (AUC) for the runs in the left two panels is only 10 times lower than that of the run in the right panel, while the correct value for the AUC is zero.

Figure 7. 2D visualization of multiple co-eluting features

Left: XIC of m/z 891.43 extracted with 30ppm tolerance. Right: 2D map of three isotopic clusters with monoisotopic peaks approximately at m/z 891.43. The XIC shows three possible elution peaks for the specified mass, however in the 2D map view it is clear that the m/z of the first (the least intense) ion is slightly shifted to lower value compared to the next two ions, which are of exactly the same mass.