Intelligent workflows for automated analysis of mass spectrometry-based proteomics data

Güler, A.T.

CHAPTER 5

5

Metadata–driven Calibration of Mass Spectrometry Data

Arzu Tugce Guler[1], Magnus Palmblad[1]

[1] Center for Proteomics and Metabolomics, Leiden University Medical Center, PO Box 9600, 2300 RC, Leiden, The Netherlands

Manuscript in preparation

Abstract

Accurate determination of ion masses by the mass spectrometer increases the confidence of identifications and eventually leads to better identification and quantification. Although mass measurement accuracy and resolving power of mass spectrometers improved significantly throughout the years, there is a certain degree of systematic and random error in every data, depending on the instrument type. It is possible and beneficial to reduce mass measurement error after mass spectrometry analysis using computational methods. Here, we present a modular, command-line tool that performs automatic internal MS1 recalibration on mzXML files. msRecal selects the suitable calibration function based on the instrument type acquired from metadata and uses the calculated exact ion masses of high confidence identifications as calibrants.

Introduction

Advances in liquid chromatography – mass spectrometry have made the high throughput analysis of proteomics data more efficient and reliable. In bottom-up analyses, samples are very complex, and many peptides can elute at the same time. Thus, achieving high mass accuracy in MS1 measurements is important since precursor mass acts as an initial filter to identify peptides[1,2]. Due to instrumental factors, there is always a degree of deviation from the exact mass in measurements, affecting the accuracy and resulting in bias. Random errors are also present in measurements, affecting the precision. Taking repeated measurements of the same sample is not a practical solution to overcome these errors, as it is usually not feasible when working with biological samples, and yet, the systematic error remains an issue to tackle in any case[3]. Calibrating measured masses with a calibration function that uses calculated exact masses, i.e., theoretical masses, as calibrants is an efficient way to reduce systematic and random error[4,5]. Typically, calibration functions take the physics of the mass analyzer into account. There are several functions available in the literature for common mass analyzer types. Choosing the correct calibration function with suitable calibrants is essential for a good calibration[6]. Getting the instrument type from the metadata and choosing the correct calibration function and parameters according to this information is useful for automating mass calibration.

Open mass spectrometry data formats such as mzXML[7] and mzML[8] usually contain metadata containing details about the instrument type. Human Proteome Organization (HUPO) Proteomics Standards Initiative's controlled vocabulary for mass spectrometry (PSI-MS CV) defines mass spectrometry-related entities in a hierarchical manner, including mass analyzer type[9]. The PSI-MS CV directly supports open formats such as mzML, mzIdentML[10], and mzTab[11]; however, their standardized annotation is not enforced in the mzXML format[12,13]. Nevertheless, it is still possible to parse relevant information from the human-readable metadata present in mzXML files.

In principle, calibrants could be chosen among the peptides already identified with high confidence in the same analysis; however, it is also possible to use the identifications from a different MS run after additional steps if the analyzed samples are very similar or the same. Palmblad et al. showed that exact masses of peptides identified by MS/MS in an ion trap instrument could be used to calibrate MS1 spectra from an FTICR instrument to reduce the overall mass measurement error after aligning the retention times[14]. Here, we focus on data from hybrid instruments, where the data is recalibrated by using peptides identified in the same MS run.

Methods

msRecal takes mzXML and pepXML[15] files as inputs and uses peptide identifications from the pepXML file to recalibrate the MS1 spectra and MS2 precursor masses in the mzXML file. The program outputs a recalibrated and reindexed mzXML file, ready to be used in different analysis pipelines. In principle, mzXML and pepXML files could be from different MS runs on similar samples. Retention times of different MS runs should be aligned before running msRecal. If identifications from the same MS run are used as calibrants, there is no need for this additional step.

msRecal is a command-line tool programmed in C. Dedicated libraries are used to read/write mzXML and pepXML files, and the GNU Scientific Library[16] is used to fit the calibration function. msRecal does not change the nature of the data; the input and output are of the same data type, mzXML. msRecal could be seamlessly incorporated into bottom-up MS analysis workflows that work with mzXML and pepXML, like the ones used by Bruin et al.[17] and Hussaarts et al.[18] An example workflow structure incorporating the msRecal module is shown in Figure 5.1. After recalibration with msRecal, the recalibrated mzXML file can be searched again with the same parameters for possible new identifications. However, since the MS2 precursor masses are updated with more accurate masses, it is also possible to do this search within a narrower error window than the initial search.

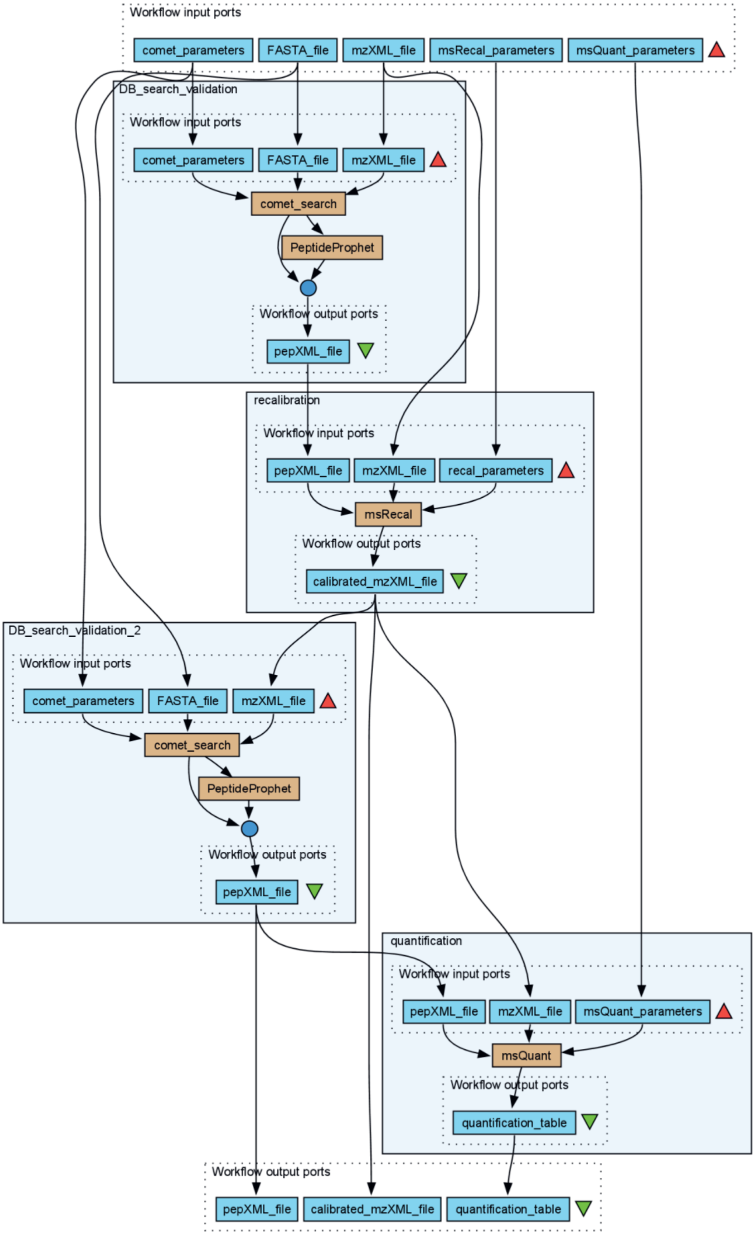

Figure 5.1. The mzXML file is recalibrated using the exact masses of the peptides from the pepXML file, obtained from the database search of the same mzXML file. The calibrated mzXML file is searched again for possible new identifications. The pepXML file with the results of the new search and the recalibrated mzXML can be used in other analysis modules downstream, e.g., in a quantification module.

The user can override the default values for parameters such as the minimum number of calibrants, maximum mass measurement error allowed for calibrants, internal mass measurement error target after calibration, threshold for background intensity, score type and threshold scores, retention time window for matching calibrants. It is highly recommended that certain parameters like ‘threshold for background intensity’, ‘maximum mass measurement error allowed’ are chosen by the user. The optimal values for these parameters vary from one data to another and may affect the calibration efficiency.

Application of the correct calibration function is the most critical step in the program. Currently, msRecal makes use of three calibration functions that are specific to instrument types[19,20,21,22].

$$\text{Orbitrap} \qquad \frac{m}{z} = \frac{A}{f^2} \qquad (1)$$

$$\text{FTICR} \qquad \frac{m}{z} = \frac{A}{f+B} \qquad (2)$$

$$\text{TOF} \qquad \frac{m}{z} = \frac{t-B}{A} \qquad (3)$$

where *A*, *B*, and *C* are the calibration coefficients; *f* is the frequency; *t* is the time.

The calibration function is chosen according to the ‘mass analyzer type’ or ‘instrument type’ parameters. Normally, these parameters are parsed from the metadata in mzXML unless the user overrides them. The PSI-MS CV defines the three mass analyzer types that the program recognizes. (Figure 5.2) It is possible that the ‘mass analyzer type’ is missing, or sometimes even incorrect, in the mzXML metadata. However, in most cases, ‘instrument type’ is given correctly. If the ‘mass analyzer type’ is missing or deemed incorrect by the program, then the ‘instrument type’ is used to set the correct value for the former. The PSI-MS CV does not define a direct relationship between the children of ‘instrument type’ and ‘mass analyzer type’ entities, so we assume a hypothetical relationship to match them, as shown in Figure 5.2.

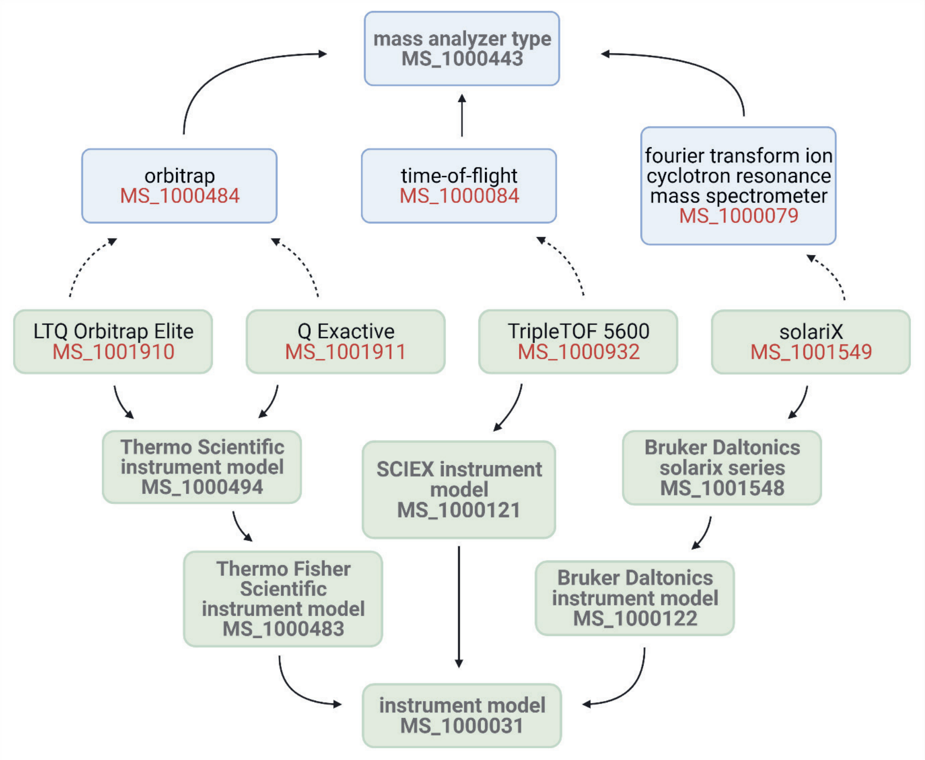

Figure 5.2. The PSI-MS controlled vocabulary groups the mass analyzer types and the instrument models separately. A direct match between the calibration function and the mass analyzer type is the most straightforward approach; however, it is also possible to make an indirect inference (shown with dashed lines) of the mass analyzer type if only the instrument model is provided. For instance, if the mass analyzer type is not given in the metadata, but the instrument model is stated as "LTQ Orbitrap Elite", then we use the function for Orbitrap.

msRecal uses the exact masses of peptides identified with high confidence to recalibrate the mass spectrometry data, thus first builds a peptide set from the pepXML file by selecting the peptides that fit the criteria, i.e., thresholds scores. In this version of msRecal, only unmodified peptides without isotope errors are used as calibrants, and the mass-to-charge ratios are calculated up to z = +4 charge state. The user could set the upper and lower score thresholds for selecting the high confidence peptides; by default, peptides with an expect score < 0.01 are selected. In addition to peptides, polydimethylcyclosiloxanes ($CH_3[Si(CH_3)_2O]_nSi(CH_3)_3$) are also added to the list of potential calibrants as they may be present when nanoelectrospray ionization is

used[23,24]. Within the matching scan/retention time window, by default [-30s,+90s], the maximum number of eligible calibrants are selected for each MS1. Only the peaks above the background intensity threshold are used, and the potential calibrants within the specified maximum mass measurement error window are matched to each peak. The suitable calibration function is used to calibrate each MS1 spectrum individually. This is done by taking the partial derivatives of the calibration function with respect to each calibration coefficient and then using the least-squares fit.
For instance, for the Orbitrap calibration function given in Eq. (1), the partial derivate with respect to its single coefficient is,

$$\frac{\partial(m/z)}{\partial A} = \frac{1}{f^2} \tag{4}$$

Next, a dummy unit for *f* is derived from the original calibration function, Eq. (1),

$$f = \frac{1}{\sqrt{m/z}} \tag{5}$$

The least-square minimization is applied first using all the measured calibrant *m/z* for an individual MS1 scan and their calculated *m/z* to find the optimal value for coefficient *A.* The calibration step is iterated several times while removing the calibrants that do not fit the function better than a given internal target, by default 2 ppm, as long as a specified minimum number of calibrants, by default 3, remain. Finally, the function in Eq. (1) is applied on the measured peak masses, using the calculated optimal value for coefficient *A,* and *f* in dummy units. Thus, the final equation used for calculating the calibrated *m/z* for an Orbitrap will be,

$$\left(\frac{m}{z}\right)' = A * \left(\frac{m}{z}\right) \tag{6}$$

where (m/z)' is the calibrated mass-to-charge ratio, *A* is the calculated calibration coefficient, and (m/z) is the measured mass-to-charge ratio.

It should be noted that the calibration coefficients of individual MS1 scans are used to calibrate MS1 peaks and the precursor masses of the corresponding MS2 scans. There is an option to exclude the uncalibrated scans in the output; otherwise, the original masses of the calibrated scans are replaced with the calibrated masses in the mzXML file while the uncalibrated scans are left as is. The file is also reindexed so that the outputted mzXML is ready to be used.

The msRecal is demonstrated on different instruments to show the calibration performance. We used publicly available data from PRIDE with accession numbers PXD000563[25] for Orbitrap, PXD000071[26] for TOF, PXD004678[27] for FTICR. The datasets come from hybrid instruments, so we used the database search results of the same data to select the calibrants. The database searches and peptide validations were performed using Comet[28] version 2021.01 rev. 0 and PeptideProphet[29], respectively, in Trans-Proteomic Pipeline[30] v6.0.0. The Homo sapiens reference proteome downloaded from Uniprot[31] on October 2021, containing 78139 entries were used in the database search. It is, of course, possible to use other database search tools and pipelines that outputs the identifications in pepXML format. We used the default expect score < 0.01 in Orbitrap data and the PeptideProphet probability matching FDR < 0.01 in TOF and FTICR data as a threshold for high confidence peptides. Mass measurement error and background thresholds are chosen based on individual data. After the calibration, the outputted mzXML is searched again with Comet using the same parameters to check the improvement in mass measurement accuracy.

Results

The mass measurement error distributions of high confidence monoisotopic peptides before and after a single calibration are shown in Figure 5.3. The same thresholds used for selecting the calibrants were applied to select the high confidence peptides.

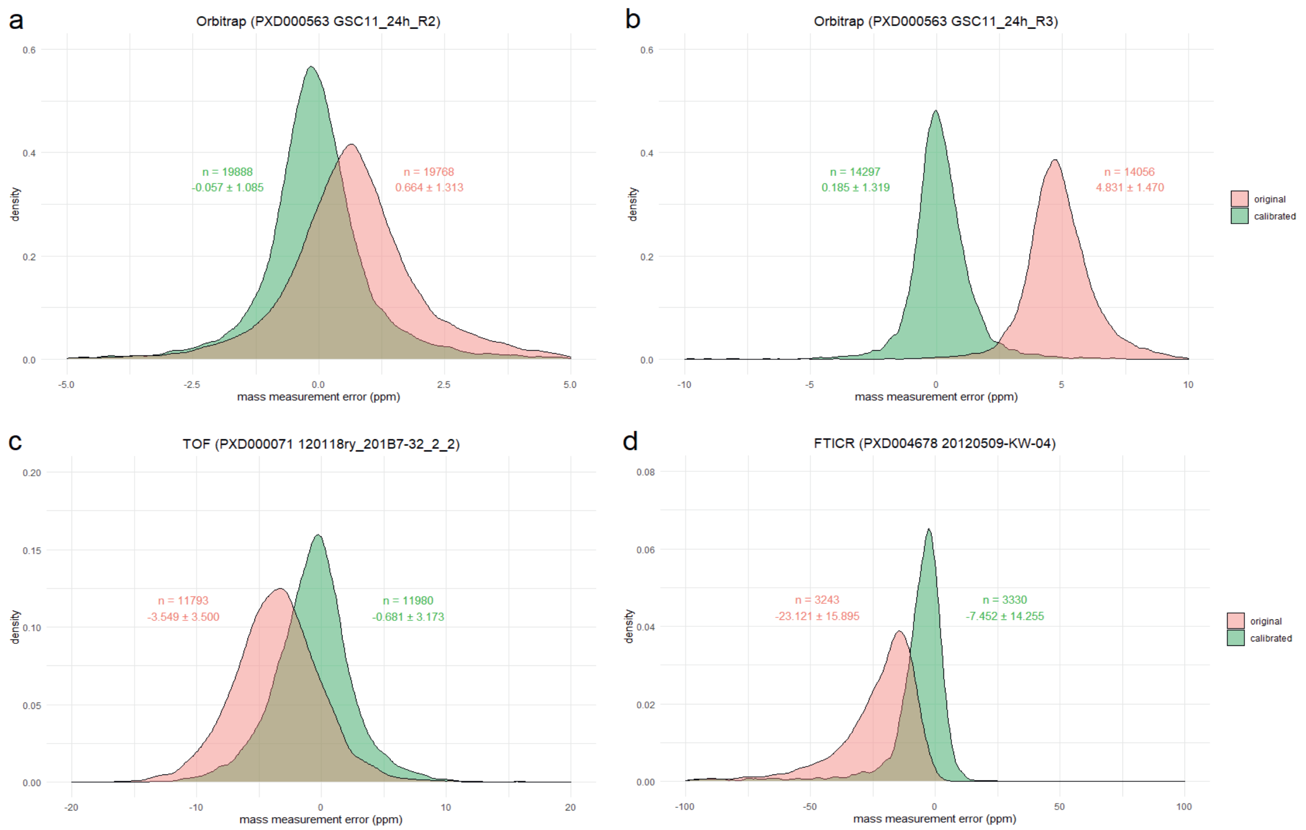

Figure 5.3. Mass measurement error distributions of original (pink) and recalibrated (green) Orbitrap (**a**, **b**), TOF (**c**) and FTICR (**d**) data. Only the high confidence peptides (expect value < 0.01 for a, b; FDR < 0.01 for c, d) without isotopic error are shown. The number of peptides, the mean, and standard deviation are also given.

As shown in figure 5.3, the mass measurement error distributions tend to center around zero and get narrower after recalibration. The mean mass measurement errors were < 1 ppm in Orbitrap and TOF data. The FTICR data already had a substantial residual bias to start with and did not have a high number of peptide identifications. Although the calibration improved both the accuracy and precision in the FTICR data, there was still some residual bias. Since the precursor masses are closer to their exact masses after calibration, searching the recalibrated data with the same parameters yielded more high confidence peptides in all.

Discussion and conclusions

The systematic and the random error decreases after recalibration, which is also the case with msRecal. Calibration performance, however, is dependent on many factors. Applying the correct calibration function is obviously the most important step, and msRecal tries to make this selection safe and automated by extracting relevant information from metadata. The number of high confidence peptides in the initial search is also a factor since having many potential calibrants increases the chances of good fits for the calibration function. On the other hand, significant mass deviation in the original data could have a negative impact on calibration performance. Even though this may already point to some issues in the original MS run, in most cases, msRecal still improves the mass error to a certain degree in such data. The improvement in peptide identifications could be observed better if the original and recalibrated data were searched in a narrower ppm range. The minimization of mass measurement error is beneficial for identification and should also improve quantification, as more peaks will be found within narrow mass measurement search windows in an MS1-based quantification. In this version of the software, only monoisotopic masses and unmodified peptides are used as calibrants. We plan to use them in future versions of the software as they could improve calibration performance in certain datasets.

The mzXML data format is still widely used, although mzML is (very) slowly replacing this format. However, since the PSI-MS CV annotation is not strictly enforced in mzXML, incomplete and even incorrect analyzer types are sometimes given in the metadata. For instance, the mass analyzer type for a QExactive instrument is annotated as a quadrupole in some datasets, whereas the mass analyzer used to acquire the data is the Orbitrap, while the quadrupole is only used as a filter for selecting the precursors. For the time being, we try to come over this issue by resorting to the instrument model information. However, in the future, with extended vendor support of PSI-MS CV terms, this could be solved more easily. Marissen and Palmblad recently published a calibration method for mzML[5]. msRecal can be seen as a complement to their work since the mzXML format is still very popular and an automated calibration tool for this data type is very useful for reanalyzing publicly available data.

The msRecal tool can be incorporated into any mass spectrometry analysis workflow that analyzes data in mzXML format, as the output format is also an mzXML file. The recalibrated mzXML can be analyzed further downstream without readjusting the existing components of the pipeline. Automatic recalibration of data in public repositories using metadata facilitates reuse of this data consistent with the FAIR principles[32].

Acknowledgments

The source code is available on GitHub https://github.com/ArzuTugceGuler/msRecal and will be maintained with bug-fixes and future updates adding more functionalities. The authors would like to thank Rob Marissen for valuable advice on the project.

References

1. Mann, M. & Kelleher, N. L. Precision proteomics: The case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 18132–18138 (2008).

2. Zubarev, R. A. & Makarov, A. Orbitrap mass spectrometry. *Anal. Chem.* **85**, 5288–5296 (2013).

3. Brenton, A. G. & Godfrey, A. R. Accurate mass measurement: Terminology and treatment of data. *J. Am. Soc. Mass Spectrom.* **21**, 1821–1835 (2010).

4. Palmblad, M., Bindschedler, L. V, Gibson, T. M., Cramer, R. & Wiley, J. Automatic internal calibration in liquid chromatography / Fourier transform ion cyclotron resonance mass spectrometry of protein digests. **20,** 3076-3080 3076–3080 (2006).

5. Marissen, R. & Palmblad, M. mzRecal: universal MS1 recalibration in mzML using identified peptides in mzIdentML as internal calibrants. *Bioinformatics* **37**, 2768-2769 (2021).

6. Romson, J. & Emmer, Å. Mass calibration options for accurate electrospray ionization mass spectrometry. *Int. J. Mass Spectrom.* **467**, 116619 (2021).

7. Pedrioli, P. G. a *et al.* A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **22**, 1459–66 (2004).

8. Martens, L. *et al.* mzML - A community standard for mass spectrometry data. *Mol. Cell. Proteomics* **10**, R110.000133 (2011).

9. Mayer, G. *et al.* The HUPO proteomics standards initiative mass spectrometry controlled vocabulary. *Database* **2013**, 1–13 (2013).

10. Jones, A. R. *et al.* The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell. Proteomics* **11**, 1–10 (2012).

11. Hoffmann, N. *et al.* MzTab-M: A Data Standard for Sharing Quantitative Results in Mass Spectrometry Metabolomics. *Anal. Chem.* **91**, 3302–3310 (2019).

12. Mayer, G. *et al.* Controlled vocabularies and ontologies in proteomics: Overview, principles and practice. *Biochim. Biophys. Acta - Proteins Proteomics* **1844**, 98–107 (2014).

13. Deutsch, E. W. *et al.* Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. *J. Proteome Res.* **16**, 4288–4298 (2017).

14. Palmblad, M. *et al.* Improving mass measurement accuracy in mass spectrometry based proteomics by combining open source tools for chromatographic alignment and internal calibration. *J. Proteomics* **72**, 722–724 (2009).

15. Keller, A., Eng, J., Zhang, N., Li, X. jun & Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **1**, 2005.0017 (2005).

16. Galassi, B. *et al. GNU Scientific Library Reference Manual - Third Edition.* (Network Theory Ltd., 2009).

17. de Bruin, J. S., Deelder, A. M. & Palmblad, M. Scientific workflow management in proteomics. *Mol. Cell. Proteomics* **11**, M111.010595 (2012).

18. Hussaarts, L. *et al.* Human Dendritic Cells with Th2-Polarizing Capacity: Analysis Using Label-Free Quantitative Proteomics. *Int. Arch. Allergy Immunol.* **174**, 170–182 (2017).

19. Christian, N. P., Arnold, R. J. & Really, J. P. Improved calibration of time-of-flight mass spectra by simplex optimization of electrostatic ion calculations. *Anal. Chem.* **72**, 3327–3337 (2000).

20. Ledford, E. B., Rempel, D. L. & Gross, M. L. Space Charge Effects in Fourier Transform Mass Spectrometry. Mass Calibration. *Anal. Chem.* **56**, 2744–2748 (1984).

21. Shi, S. D. H., Drader, J. J., Freitas, M. A., Hendrickson, C. L. & Marshall, A. G. Comparison and interconversion of the two most common frequency-to-mass calibration functions for Fourier transform ion cyclotron resonance mass spectrometry. *Int. J. Mass Spectrom.* **195–196**, 591–598 (2000).

22. Gorshkov, M. V., Good, D. M., Lyutvinskiy, Y., Yang, H. & Zubarev, R. A. Calibration function for the orbitrap FTMS accounting for the space charge effect. *J. Am. Soc. Mass Spectrom.* **21**, 1846–1851 (2010).

23. Schlosser, A. & Volkmer-Engert, R. Volatile polydimethylcyclosiloxanes in the ambient laboratory air identified as source of extreme background signals in nanoelectrospray mass spectrometry. *J. Mass Spectrom.* **38**, 523–525 (2003).

24. Haas, W. *et al.* Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Mol. Cell. Proteomics* **5**, 1326–1337 (2006).

25. Lichti, C. F. *et al.* Integrated chromosome 19 transcriptomic and proteomic data sets derived from glioma cancer stem-cell lines. *J. Proteome Res.* **13**, 191–199 (2014).

26. Yamana, R. *et al.* Rapid and deep profiling of human induced pluripotent stem cell proteome by one-shot NanoLC-MS/MS analysis with meter-scale monolithic silica columns. *J. Proteome Res.* **12**, 214–221 (2013).

27. Worah, K. *et al.* Proteomics of Human Dendritic Cell Subsets Reveals Subset-Specific Surface Markers and Differential Inflammasome Function. *Cell Rep.* **16**, 2953–2966 (2016).

28. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: An open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).

29. 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**, 5383–92 (2002).

30. Deutsch, E. W. *et al.* Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics - Clin. Appl.* **9**, 745–754 (2015).

31. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

32. Wilkinson, M. D. *et al.* Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9 (2016).