Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

What computational non-targeted mass spectrometry-based metabolomics can gain from shotgun proteomics

Hamid Hamzeiy and Jürgen Cox

Computational workflows for mass spectrometry-based shotgun proteomics and untargeted metabolomics share many steps. Despite the similarities, untargeted metabolomics is lagging behind in terms of reliable fully automated quantitative data analysis. We argue that metabolomics will strongly benefit from the adaptation of successful automated proteomics workflows to metabolomics. MaxQuant is a popular platform for proteomics data analysis and is widely considered to be superior in achieving high precursor mass accuracies through advanced nonlinear recalibration, usually leading to five to ten-fold better accuracy in complex LC–MS/MS runs. This translates to a sharp decrease in the number of peptide candidates per measured feature, thereby strongly improving the coverage of identified peptides. We argue that similar strategies can be applied to untargeted metabolomics, leading to equivalent improvements in metabolite identification.

Address
Computational Systems Biochemistry, Max-Planck Institute of Biochemistry, Martinsried, Germany

Corresponding author: Cox, Jürgen (cox@biochem.mpg.de)

Current Opinion in Biotechnology 2017, **43**:141–146

This review comes from a themed issue on **Analytical biotechnology**

Edited by **Jurre J Kamphorst** and **Ian A Lewis**

For a complete overview see the Issue and the Editorial

Available online 28th December 2016

http://dx.doi.org/10.1016/j.copbio.2016.11.014

Introduction

Mass spectrometry-based proteomics [1,2,3••] has matured during recent years to a degree that makes it readily usable as a standard research tool in many branches of biological and biomedical research. Most often proteomics is implemented in the form of shotgun proteomics, in which proteins are first digested to peptides, separated by liquid chromatography, and finally studied in the mass spectrometer as intact peptides as well as their fragmentation patterns (LC–MS/MS). Proteomics applications include expression proteomics, analysis of protein–protein interactions [4], study of post-translational modifications [5], as well as determination of subcellular localization [6], which can all be done in a dynamic, time-dependent manner [7]. Most proteomics experiments can be performed without the use of labels, owing to appropriate algorithms for relative label-free quantification [8]. The complete yeast proteome can be quantified nowadays with moderate effort and studied in many different conditions [9–11], while in human cellular proteomes a depth of 10,000 proteins can be achieved [12–15].

Ten years ago, the situation was very different. Proteomics projects were very time consuming since the data analysis was mostly done in a semi-manual fashion. While peptide database search engines [16–20] and other software and algorithms for the identification and quantification of peptides already existed in principle, a lot of manual validation was still necessary in order to obtain reliable results that could be used for solid biological interpretation.

Certainly, technological improvements like the introduction of the Orbitrap mass spectrometer [21–23] and improvements in sample preparation also contributed to today's proteomics workflows to be evermore robust and easy to use. However, a large part of the improved situation is owed to the software platforms and computational workflows that have become mature and reliable. This starts with basic activities such as feature detection, correct label assignment, and processing of MS/MS spectra. Then, the identification process can reliably be controlled by false discovery rates on the peptide-spectrum match (PSM) or protein level. Furthermore, the results of quantification methods became better than what could be achieved with manual analysis. All these improvements together lead to a situation in which shotgun proteomics data analysis is approaching a state of maturity that is comparable to next generation sequencing data analysis. Also, software tools that aid in the biological interpretation of quantitative proteomics results are available and well accepted in the community [24].

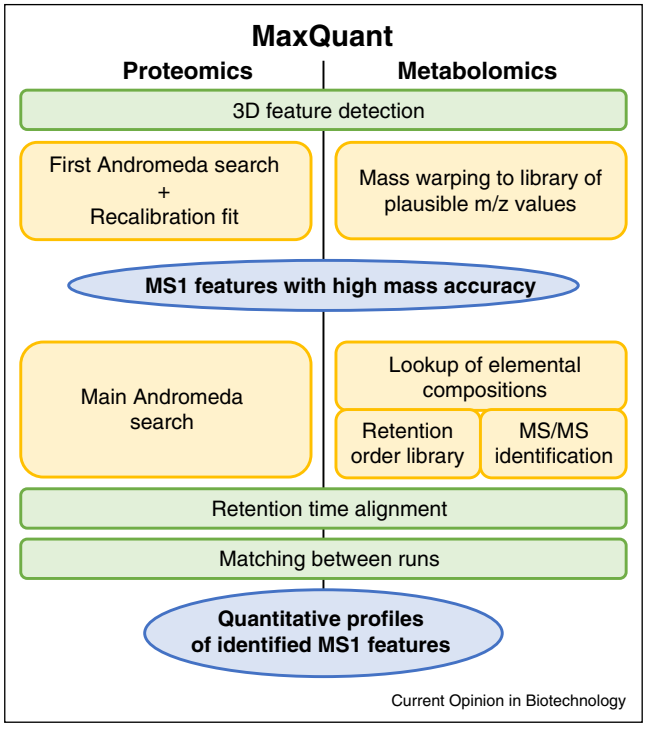
One of these computational workflows is MaxQuant [25,26•], including the Andromeda peptide search engine [27], which provides a complete solution for most standard quantitative experimental designs in shotgun proteomics. Its development provided seminal contributions to the reliable automation of the data analysis workflow. One aspect in which MaxQuant is unique is how it improves the mass accuracy of peptide features using computational techniques [28,29]. Nonlinear mass recalibration is applied to the MS1 features in an m/z and retention time dependent way. Multiple mass measurements over

elution profiles and isotopic peaks are then integrated, achieving mass accuracies in the ppb range for standard Orbitrap data in a complex proteomics run, which is a 5–10-fold increase over standard techniques.

Untargeted metabolomics [30,31] is a highly evolved field with many applications already accessible and high promises for the future. A wealth of analytical techniques [32] exist for its study and many computational tools [33,34] have been developed within the community. However, interpreting mass spectrometry-based untargeted metabolomics data remains a challenge and limits the translation of results into biologically relevant conclusions [35••]. Although the power of untargeted profiling is undeniable, it is the case that most mechanistic links are still revealed by hypothesis-driven targeted methods [36•]. This is likely due to untargeted metabolomics typically yielding complex data patterns that are not easily amenable to intuitive interpretation [36•]. One could make the provocative statement that untargeted metabolomics is several years behind shotgun proteomics in terms of ease of data analysis and interpretation.

Our plan is to create a version of MaxQuant for the analysis of untargeted metabolomics LC–MS data whose workflow follows loosely the shotgun proteomics workflow as sketched in Figure 1. Several important data processing steps can be transferred to metabolomics with only minor adaptations, as for instance the 3D feature detection, retention time alignment and matching of features between LC–MS runs based on accurate masses and retention times. Other processing steps need more changes in order to become applicable to metabolomics data. For instance, the two-level peptide search strategy, in which the identifications of the first round of searches are used to determine multidimensional nonlinear recalibration curves need to be replaced by another two-level strategy based on mass warping due to the absence of a universal search engine approach in metabolomics. We strongly believe that the application of this sort of nonlinear mass recalibration to metabolomics data is highly beneficial for compound identification by increasing the range of molecules for which an elemental composition can be assigned.

Figure 1



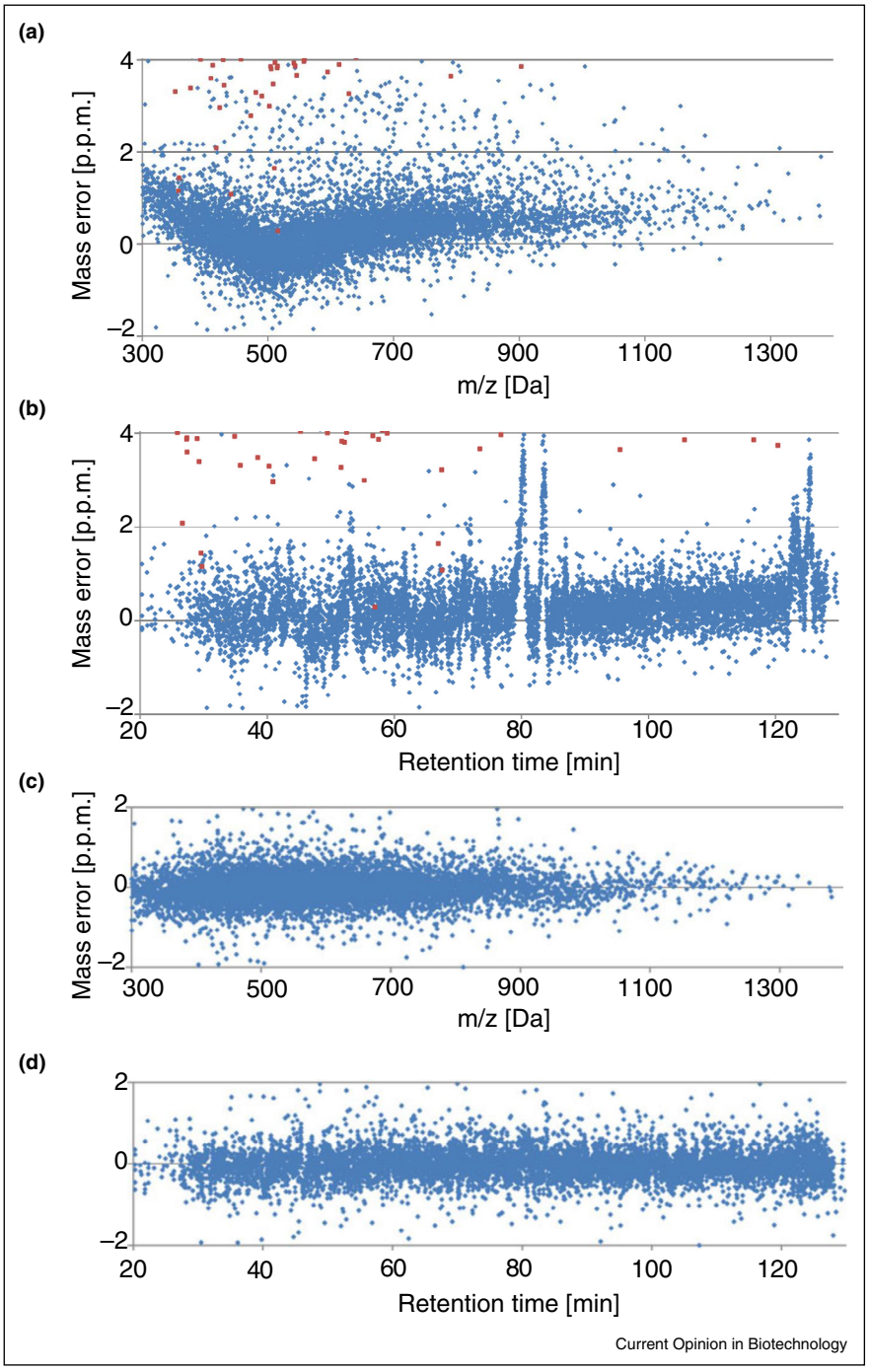
Schematic overview of high mass accuracy feature identification and quantification workflows in MaxQuant for shotgun proteomics and for untargeted metabolomics.

Improvement of mass accuracy in proteomics

Here we describe how high mass accuracy is achieved by mass recalibration algorithms in proteomics. In the next section we sketch our path to implementing similar improvements in untargeted metabolomics. For the determination of the nonlinear mass recalibration curves in proteomics we follow a strategy employing two consecutive peptide database searches (Figure 1). After having performed the 3D feature detection, a first round of Andromeda searches is performed. The purpose of this search is to generate a list of features with known masses which can then be used for recalibration. The precursor mass tolerance for the first search is relatively large, for example, 20 ppm, to be able to also correct for larger instrumental drift. Since there are many peptides available in a complex shotgun proteomics run, we can be restrictive at this stage and accept only identifications that are correct with high certainty, for example, by requiring a high Andromeda score threshold, which will typically still result in thousands of peptides per LC–MS run. Alternatively, one can use standards instead of the first search identifications. However, this strategy has the disadvantage that only a few features of known mass are available, which is usually not sufficient to perform the nonlinear recalibration to the accuracy attainable through the approach using many peptides from the sample. Figure 2(a) shows the mass deviations in a typical LC–MS run as a function of m/z, while in Figure 2(b) they are shown as a function of retention time. Clearly, just linear recalibration would leave many mass deviations far above 1 ppm.

Once we have obtained the long list of known masses, we fit a model to the mass deviations describing them by nonlinear dependencies on m/z and retention time. For time-of-flight mass spectrometers, the intensity dependence of the mass error is estimated and corrected (not necessary for Orbitrap data). No particular functional form of these dependencies is assumed. Instead, we use either splines or piecewise linear functions as models for the m/z and retention time dependencies. Overfitting is avoided

Figure 2

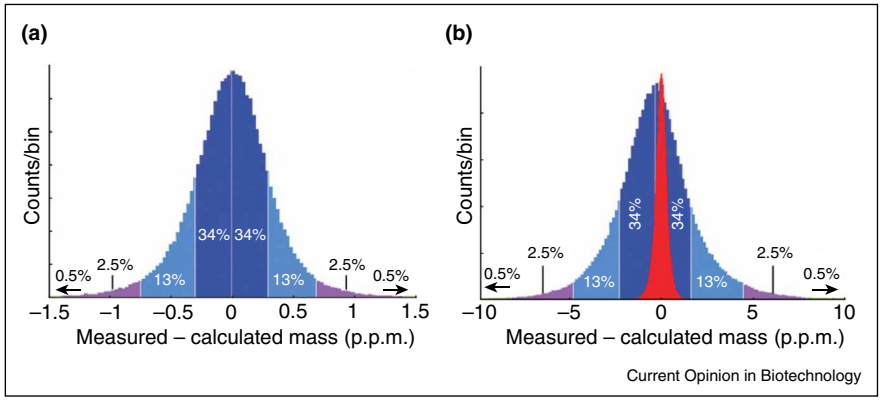


Nonlinear mass recalibration in MaxQuant. **(a)** *m/z* dependence of the mass error before recalibration on an Orbitrap mass spectrometer. **(b)** Retention time dependence of the mass error before recalibration on an Orbitrap mass spectrometer. **(c)** and **(d)** Same as **(a)** and **(b)** after application of the nonlinear mass recalibration functions.
Adapted from Ref. [28].

by keeping the ratio of number of parameters to number of data points to a low percentage number. Figure 2(c–d) shows the residuals after applying the calibration functions to the data which fluctuate independently around zero.

After having obtained the nonlinear recalibration functions, these are applied to all peptide features, also to the ones that were not used in the fit. This includes those MS1 feature that were fragmented, but not included in the

Figure 3



Mass error distributions before and after nonlinear mass recalibration. The red histogram on the right side is the same as the histogram on the left side and was added for comparison.
Adapted from Ref. [25].

recalibration fit due to the Andromeda score threshold. It also includes the MS1 features that were not fragmented at all, which is usually a vast majority of the signals [37]. Figure 3(a) shows a histogram of mass deviations obtained after recalibration in a typical LC–MS run. The average absolute mass deviation (absolute value of the difference between measured and calculated masses) is below 300 ppb. In Figure 3(b) the same histogram is recorded for masses taken directly from the instrument software without applying MaxQuant recalibration. Since the histogram is centered near zero, a linear shift as recalibration would obviously not improve the mass accuracy much. For comparison purposes, the red histogram was included which is the same as in Figure 3(a). For this typical LC–MS run the mass accuracy was improved by about 6-fold using MaxQuant recalibration routines.

Such a strong increase in mass accuracy will have implications on the peptide identification process. When searching in the human proteome the corresponding shrinkage of the precursor mass tolerance window — which is individualized for each peptide in MaxQuant — will translate proportionally into a restriction of possible peptide candidates for a given MS1 feature. Therefore, less information needs to come from the MS/MS spectrum to have the same certainty of identification of a peptide. With a fixed false discovery rate of, for example, 1% for PSMs (and 1% for proteins) the coverage of identified proteins will rise [29]. The extent of the improvement depends on many factors, like size of the protein sequence space used for generating the in silico peptide list for the database search, the type of digestion and the number of variable modifications.

Improvement of mass accuracy in metabolomics

Similar concepts as described in the previous section can be applied to non-targeted metabolomics. While our work in proteomics is mostly agnostic of the mass spectrometric instrument, here we focus on the Orbitrap since the scaling of resolution with the mass range is favorable for small masses. A central part of the proteomics workflow is the generation of MS1 features with known masses through the 'first Andromeda search' (Figure 1). In principle, one could follow a similar route and replace the peptide database search engine with a spectral library search and accept only indisputable identifications. However, we decided to adapt a different strategy that would also be applicable in the absence of MS/MS data.

We first generate a library of 'plausible m/z values' that one is likely to find in a metabolomics LC–MS run. This is in the first instance filled with all molecules from databases of compounds with biological relevance, such as CheBI [38]. Then we perform the MaxQuant 3D feature extraction on a large amount of untargeted metabolomics LC–MS runs in order to find which of the features can be interpreted as an adduct of a molecule that is already in the library of plausible m/z values, which are then also added to the library of 'plausible m/z values. The library contains all isotopic peaks, not only monoisotopic masses, since the subsequent algorithms will work on the 3D peak features before assembling them to isotope patterns.

Each LC–MS run to be analyzed is then mass aligned to this list of plausible m/z values. For this we use a kind of warping algorithm that finds an optimal nonlinear calibration function under the objectives as bringing as many MS1 features as possible as close as possible to a value in the list of plausible masses. This is done while requiring smoothness of the recalibration function in order to avoid overfitting. In this optimization procedure most of the MS1 features will 'snap' to the correct elemental composition. Some will not, because the correct composition is

not present yet in the library. The algorithm will still be able to find a good interpolating solution due to the smoothness requirement.

The library is a dynamic entity which will be updated based on the knowledge gain resulting from each alignment with an LC–MS run. If, after a new alignment there are unmatched MS1 features left with good signal-to-noise, and fitting a plausible new elemental composition, it will be added to the library. The alternative to this procedure would be to work in the space of all theoretically possible elemental compositions. However, we think there are big advantages to build up this reference list bottom up from real data and not have it filled up from the beginning with things that will never be seen in actual LC–MS runs.

The degree to which mass accuracy helps in reducing the number of possible molecular formulas depends on many factors, including the molecular mass and assumptions on the space of possible formulas. Under reasonable assumptions the number of candidate formulas shrinks considerably when going from 5 ppm to sub ppm accuracy over a wide range of masses as shown in Table 3 of Ref. [39]. Orthogonal filters like isotopic abundance ratios or ion mobility measurements would certainly diminish the number of candidates as well. Preliminary results show that the increase in mass accuracy obtained by our proposed method is indeed comparable to the gains seen in shotgun proteomics. The resulting reduction in candidates will lead to complete determination of elemental compositions for the majority of MS1 features. This will improve MS1-only workflows that use a lab-specific retention order library for distinguishing isomers. Metabolic flux [40–43] analysis can be supported as well by including the ^{13}C-labeling patterns of metabolic intermediates or end products into the list of plausible m/z values.

Conclusions

The adaptation of MaxQuant to untargeted metabolomics will strongly improve the mass accuracy of MS1 features. Similar to proteomics, this increased identification information will strengthen the robustness of the automated data analysis workflow in untargeted metabolomics. Together with other features from the MaxQuant workflow that are readily transferable to metabolomics — retention time alignment and matching between runs — MaxQuant should yield a useful addition to the computational metabolomics toolbox.

Conflict of interest statement

The authors declare no competing financial interests.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 686547.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Cox J, Mann M: **Quantitative, high-resolution proteomics for data-driven systems biology**. *Annu Rev Biochem* 2011, **80**:273-299.

2. Altelaar AF, Munoz J, Heck AJ: **Next-generation proteomics: towards an integrative view of proteome dynamics**. *Nat Rev Genet* 2013, **14**:35-48.

3. •• Aebersold R, Mann M: **Mass-spectrometric exploration of proteome structure and function**. *Nature* 2016, **537**:347-355.
This is a recent review of mass spectrometry-based proteomics summarizing its achievements and the remaining challenges. It is summarized how mass-spectrometry-based proteomics has matured from a largely technology-driven field of research into a mainstream analytical tool for the life sciences.

4. Hein MY, Hubner NC, Poser I, Cox J, Nagaraj N, Toyoda Y, Gak IA, Weisswange I, Mansfeld J, Buchholz F *et al.*: **A human interactome in three quantitative dimensions organized by stoichiometries and abundances**. *Cell* 2015, **163**:712-723.

5. Sharma K, D'Souza RC, Tyanova S, Schaab C, Wisniewski JR, Cox J, Mann M: **Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling**. *Cell Rep* 2014, **8**:1583-1594.

6. Itzhak DN, Tyanova S, Cox J, Borner GH: **Global, quantitative and dynamic mapping of protein subcellular localization**. *Elife* 2016:5.

7. Robles MS, Cox J, Mann M: **In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism**. *PLoS Genet* 2014, **10**:e1004047.

8. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M: **Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ**. *Mol Cell Proteomics* 2014, **13**:2513-2526.

9. Hebert AS, Richards AL, Bailey DJ, Ulbrich A, Coughlin EE, Westphall MS, Coon JJ: **The one hour yeast proteome**. *Mol Cell Proteomics* 2014, **13**:339-347.

10. de Godoy LM, Olsen JV, Cox J, Nielsen ML, Hubner NC, Frohlich F, Walther TC, Mann M: **Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast**. *Nature* 2008, **455**:1251-1254.

11. Stefely JA, Kwiecien NW, Freiberger EC, Richards AL, Jochem A, Rush MJ, Ulbrich A, Robinson KP, Hutchins PD, Veling MT *et al.*: **Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling**. *Nat Biotechnol* 2016.

12. Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, Paabo S, Mann M: **Deep proteome and transcriptome mapping of a human cancer cell line**. *Mol Syst Biol* 2011, **7**:548.

13. Beck M, Schmidt A, Malmstroem J, Claassen M, Ori A, Szymborska A, Herzog F, Rinner O, Ellenberg J, Aebersold R: **The quantitative proteome of a human cell line**. *Mol Syst Biol* 2011, **7**:549.

14. Munoz J, Low TY, Kok YJ, Chin A, Frese CK, Ding V, Choo A, Heck AJ: **The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells**. *Mol Syst Biol* 2011, **7**:550.

15. Mann M, Kulak NA, Nagaraj N, Cox J: **The coming age of complete, accurate, and ubiquitous proteomes**. *Mol Cell* 2013, **49**:583-590.

16. Eng JK, McCormack AL, Yates JR: **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database**. *J Am Soc Mass Spectrom* 1994, **5**:976-989.

17. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data**. *Electrophoresis* 1999, **20**:3551-3567.

18. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: **Open mass spectrometry search algorithm**. *J Proteome Res* 2004, **3**:958-964.

19. Craig R, Beavis RC: **TANDEM: matching proteins with tandem mass spectra**. *Bioinformatics* 2004, **20**:1466-1467.

20. Bern M, Cai Y, Goldberg D: **Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry**. *Anal Chem* 2007, **79**:1393-1400.

21. Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R: **The Orbitrap: a new mass spectrometer**. *J Mass Spectrom* 2005, **40**:430-443.

22. Olsen JV, de Godoy LM, Li G, Macek B, Mortensen P, Pesch R, Makarov A, Lange O, Horning S, Mann M: **Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap**. *Mol Cell Proteomics* 2005, **4**:2010-2021.

23. Michalski A, Damoc E, Hauschild JP, Lange O, Wieghaus A, Makarov A, Nagaraj N, Cox J, Mann M, Horning S: **Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer**. *Mol Cell Proteomics* 2011, **10** M111.011015.

24. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J: **The perseus computational platform for comprehensive analysis of (prote)omics data**. *Nat Methods* 2016, **13**:731-740.

25. 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, **26**:1367-1372.

26. • Tyanova S, Temu T, Cox J: **The MaxQuant computational platform for mass spectrometry-based shotgun proteomics**. *Nat Protoc* 2016, **11**:2301-2319.

This is a protocol describing the usage of MaxQuant for shotgun proteomics data analysis on a large variety of experimental designs and quantification strategies.

27. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M: **Andromeda: a peptide search engine integrated into the MaxQuant environment**. *J Proteome Res* 2011, **10**:1794-1805.

28. Cox J, Michalski A, Mann M: **Software lock mass by two-dimensional minimization of peptide mass errors**. *J Am Soc Mass Spectrom* 2011, **22**:1373-1380.

29. Cox J, Mann M: **Computational principles of determining and improving mass precision and accuracy for proteome measurements in an Orbitrap**. *J Am Soc Mass Spectrom* 2009, **20**:1477-1485.

30. Patti GJ, Yanes O, Siuzdak G: **Innovation: metabolomics: the apogee of the omics trilogy**. *Nat Rev Mol Cell Biol* 2012, **13**:263-269.

31. Fuhrer T, Zamboni N: **High-throughput discovery metabolomics**. *Curr Opin Biotechnol* 2015, **31**:73-78.

32. Zhang A, Sun H, Wang P, Han Y, Wang X: **Modern analytical techniques in metabolomics analysis**. *Analyst* 2012, **137**:293-300.

33. Misra BB, van der Hooft JJ: **Updates in metabolomics tools and resources: 2014–2015**. *Electrophoresis* 2016, **37**:86-110.

34. Alonso A, Marsal S, Julia A: **Analytical methods in untargeted metabolomics: state of the art in 2015**. *Front Bioeng Biotechnol* 2015, **3**:23.

35. •• 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.

In this publication emerging technologies that can be applied after untargeted profiling to extend biological interpretation of metabolomic data are reviewed. Recent advances are highlighted that help transform untargeted profiling results into structures, concentrations, pathway fluxes and localization patterns.

36. • Sevin DC, Kuehne A, Zamboni N, Sauer U: **Biological insights through nontargeted metabolomics**. *Curr Opin Biotechnol* 2015, **34**:1-8.

The authors compare the contributions of traditional targeted and non-targeted metabolomics in advancing different research areas. They conclude that novel computational approaches are required to tap the full potential of nontargeted metabolomics.

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

38. de Matos P, Alcantara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C: **Chemical entities of biological interest: an update**. *Nucleic Acids Res* 2010, **38**:D249-D254.

39. Kind T, Fiehn O: **Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm**. *BMC Bioinform* 2006, **7**:234.

40. Zamboni N: **^{13}C metabolic flux analysis in complex systems**. *Curr Opin Biotechnol* 2011, **22**:103-108.

41. Munger J, Bennett BD, Parikh A, Feng XJ, McArdle J, Rabitz HA, Shenk T, Rabinowitz JD: **Systems-level metabolic flux profiling identifies fatty acid synthesis as a target for antiviral therapy**. *Nat Biotechnol* 2008, **26**:1179-1186.

42. Yuan J, Bennett BD, Rabinowitz JD: **Kinetic flux profiling for quantitation of cellular metabolic fluxes**. *Nat Protoc* 2008, **3**:1328-1340.

43. Wiechert W, Noh K: **Isotopically non-stationary metabolic flux analysis: complex yet highly informative**. *Curr Opin Biotechnol* 2013, **24**:979-986.