Tools for Label-free Peptide Quantification*[S]

Sven Nahnsen‡‡, Chris Bielow§, Knut Reinert§, and Oliver Kohlbacher‡¶

The increasing scale and complexity of quantitative proteomics studies complicate subsequent analysis of the acquired data. Untargeted label-free quantification, based either on feature intensities or on spectral counting, is a method that scales particularly well with respect to the number of samples. It is thus an excellent alternative to labeling techniques. In order to profit from this scalability, however, data analysis has to cope with large amounts of data, process them automatically, and do a thorough statistical analysis in order to achieve reliable results. We review the state of the art with respect to computational tools for label-free quantification in untargeted proteomics. The two fundamental approaches are feature-based quantification, relying on the summed-up mass spectrometric intensity of peptides, and spectral counting, which relies on the number of MS/MS spectra acquired for a certain protein. We review the current algorithmic approaches underlying some widely used software packages and briefly discuss the statistical strategies for analyzing the data. ***Molecular & Cellular Proteomics 12: 10.1074/mcp.R112.025163, 549–556, 2013.***

Over recent decades, mass spectrometry has become the analytical method of choice in most proteomics studies (*e.g.* Refs. 1–4). A standard mass spectrometric workflow allows for both protein identification and protein quantification (5) in some form. For a long time, the technology has been used mainly for qualitative assessments of protein mixtures, namely, to assess whether a specific protein is in the sample or not. However, for the majority of interesting research questions, especially in the field of systems biology, this binary information (present or not) is not sufficient (6). The necessity of more detailed information on protein expression levels drives the field of quantitative proteomics (7, 8), which enables the integration of proteomics data with other data sources and allows network-centered studies, as reviewed in Ref. 9. Recent studies show that mass-spectrometry-based quantitative proteomics experiments can provide quantitative information (relative or absolute) for large parts, if not the entire set, of expressed proteins (10–12).

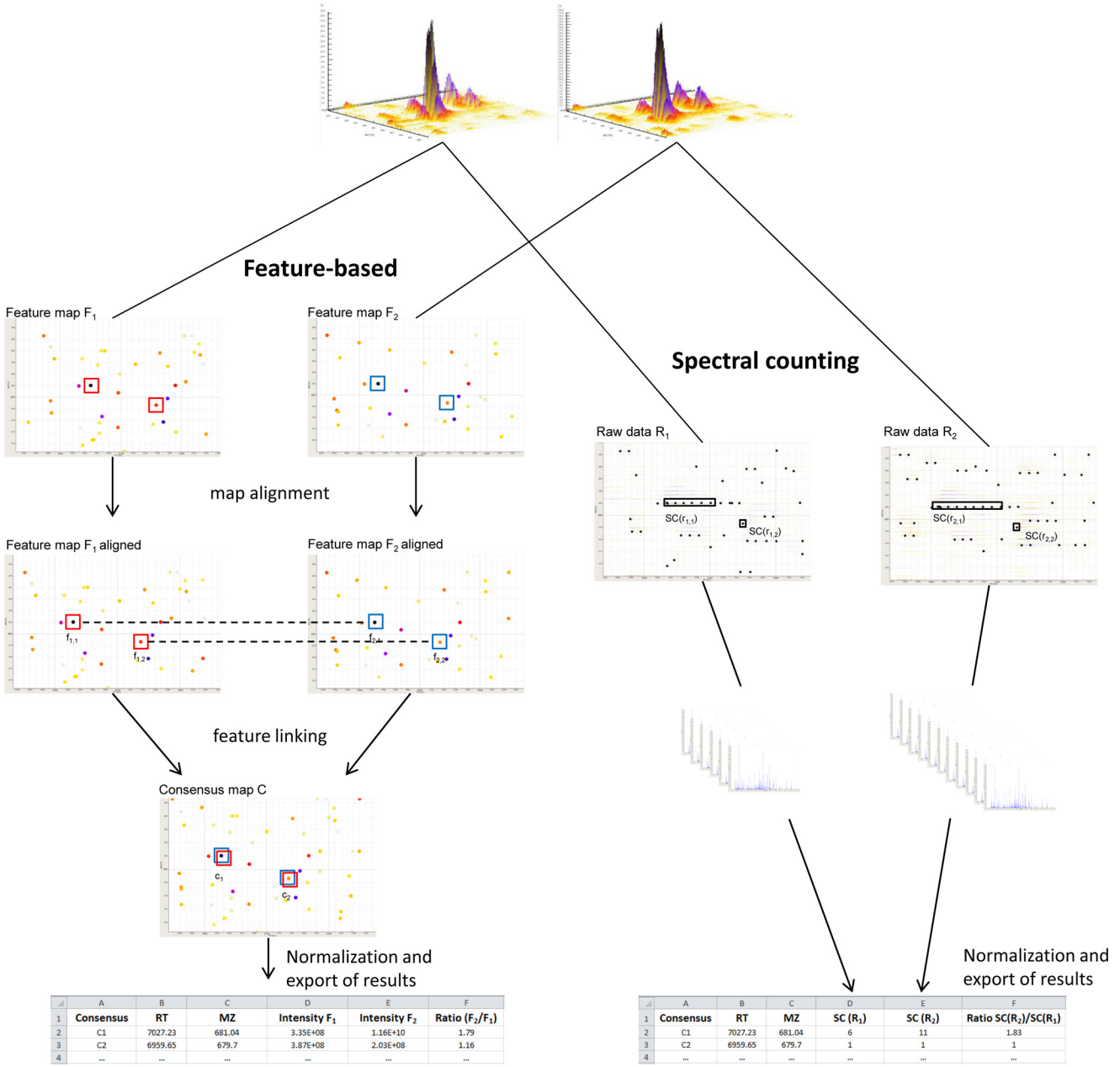
From the ‡Center for Bioinformatics, Quantitative Biology Center and Department of Computer Science, University of Tübingen, Sand 14, 72076 Tübingen, Germany; §Institute of Computer Science, Freie Universität, 14195 Berlin, Germany; ‡‡Quantitative Biology Center, University of Tübingen, 72076 Tübingen, Germany

Received October 23, 2012, and in revised form, October 23, 2012
Published, MCP Papers in Press, December 17, 2012, DOI 10.1074/mcp.R112.025163

Since the isotope-coded affinity tag protocol was first published in 1999 (13), numerous labeling strategies have found their way into the field of quantitative proteomics (14). These include isotope-coded protein labeling (15), metabolic labeling (16, 17), and isobaric tags (18, 19). Comprehensive overviews of different quantification strategies can be found in Refs. 20 and 21. Because of the shortcomings of labeling strategies, label-free methods are increasingly gaining the interest of proteomics researchers (22, 23). In label-free quantification, no label is introduced to either of the samples. All samples are analyzed in separate LC/MS experiments, and the individual peptide properties of the individual measurements are then compared. Regardless of the quantification strategy, computational approaches for data analyses have become the critical final step of the proteomics workflow. Overviews of existing computational approaches in proteomics are provided in Refs. 24 and 25. The computational label-free quantification workflow in visualized in Fig. 1. Comparing peptide quantities using mass spectrometry remains a difficult task, because mass spectrometers have different response values for different chemical entities, and thus a direct comparison of different peptides is not possible. The computational analysis of a label-free quantitative data set consists of several steps that are mainly split in raw data signal processing and quantification. Signal processing steps comprise data reduction procedures such as baseline removal, denoising, and centroiding.

These steps can be accomplished in modular building blocks, or the entire analysis can be performed using monolithic analysis software. Recently, it has been shown that it is beneficial to combine modular blocks from different software tools to a consensus pipeline (26). The same study also illustrates the diversity of methods that are modularized by different software tools. In another recent publication, monolithic software packages are compared (27). In that study, the authors identify a set of seven metrics: detection sensitivity, detection consistency, intensity consistency, intensity accuracy, detection accuracy, statistical capability, and quantification accuracy. Despite the missing independence of these metrics and the loose reporting of software parameter settings, such comparative studies are of great interest to the field of quantitative proteomics. A general conclusion from these studies is that the choice of software might, to a certain degree, affect the final results of the study.

Absolute quantification of peptides and proteins using intensity-based label-free methods is possible and can be done with excellent accuracy, if standard addition is used. With the



	A	B	C	D	E	F
1	Consensus	RT	MZ	Intensity F_1	Intensity F_2	Ratio (F_2/F_1)
2	C1	7027.23	681.04	3.35E+08	1.16E+10	1.79
3	C2	6959.65	679.7	3.87E+08	2.03E+08	1.16
4	...	...	...	...	...	...

	A	B	C	D	E	F
1	Consensus	RT	MZ	SC (R_1)	SC (R_2)	Ratio SC(R_2)/SC(R_1)
2	C1	7027.23	681.04	6	11	1.83
3	C2	6959.65	679.7	1	1	1
4	...	...	...	...	...	...

FIG. 1. **The sample cohort that can be analyzed via label-free proteomics is not limited in size.** Each sample is processed separately through the sample preparation and data acquisition pipeline. For data analysis, the data from the different LC/MS runs are combined.

help of known concentrations, calibration lines can be drawn, and absolute protein quantities can be directly inferred from these calibration measurements (28). Furthermore, it has been suggested that peptide peak intensities can be predicted and absolute quantities can be derived from these predictions (29). However, the limited accuracy of predictions or the need for peptides of known concentrations limits these approaches to selected proteins/peptides only and prevents their use on a proteome-wide scale.

Spectral counting methods have also been used for the estimation of absolute concentrations on a global scale (30), albeit at drastically reduced accuracy relative to intensity-based methods. In one study, the authors used a mixture of 48 proteins with known concentrations and predicted the absolute copy number amounts of thousands of proteins based on that mixture. Despite the fact that large, proteome-wide data sets will dilute the effects of different peptide detectabilities on the individual protein level, such methods will always be limited in their accuracy of quantification.

The generic nature of label-free quantification is not restricted to any model system and can also be employed with tissue or body fluids (31, 32). However, the label-free ap-

proach is more sensitive to technical deviations between LC/MS runs as information is compared between different measurements. Therefore, the reproducibility of the analytical platform is crucial for successful label-free quantification. The recent success of label-free quantification could only be accomplished through significant improvements of algorithms (33–36). An increasingly large collection of software tools for label-free proteomics have been published as open source applications or have entered the market as commercially available packages. This review aims at outlining the computational methods that are generally implemented by these software tools. Furthermore, we illustrate strengths and weaknesses of different tools. The review provides an information resource for the broad proteomics audience and does not illustrate all algorithmic details of the individual tools.

MATERIALS AND METHODS

The Nature of LC-MS/MS Data—Quantitative proteomics data from LC/MS and/or LC-MS/MS experiments typically have a large data volume (tens to hundreds of gigabytes per sample are not uncommon), and the data are rather complex. Typically, digested proteins (*i.e.* complex peptide samples) are separated on a liquid chromatography column and ionized, and the resulting MS spectra are recorded by a mass spectrometer. For MS/MS experiments, peptide ions are selected (based on their intensity or through an inclusion list) for fragmentation and fragment ion spectra are recorded. These MS/MS spectra usually form the basis of the identification (which we do not consider here), but they also can be used for spectral counting.

Depending on the resolution of the mass analyzer, and because the ionization is a stochastic process, even identical ions will not be measured at the exact same *m/z*; instead they form a distribution of measurements around the true *m/z* value. This distribution is called a (raw) peak and can be described by a mathematical model (a normal distribution is a good approximation, but not quite sufficient). The process of *peak picking* or *centroiding* aims at estimating the parameters of the peak model, such as the *centroid*, *intensity*, *width*, and *skew*. Centroiding reduces the raw measurement data to a handful of parameters for each compound and, most important, yields a single value for the *m/z* of the ion. The centroid *m/z* can be reported as the position of the maximal intensity, or by averaging over *m/z* (raw data points weighted by intensity). Likewise, the intensity of a peak can be read off as the maximum height from the raw data (the *peak apex*), or one can compute the area under the curve (*i.e.* the peak volume). It is important to know whether the data are centroided or not, because some software can handle only one type of input data.

Fig. 1 shows a typical data set generated from a biological sample using HPLC-MS and illustrates its multidimensional nature. After being eluted from the column, analytes are continuously injected into the mass spectrometer, which records mass spectra (scans) at high speed. Stacking individual spectra yields a three-dimensional dataset, a so-called map. When peptide mass spectrometry is preceded by liquid chromatography fractionation, the observed signal corresponding to a single charge state of a peptide is actually a two-dimensional intensity distribution in retention time and mass-to-charge The data points belonging to this distribution are called a *feature* (*e.g.* the two-dimensional signal in Fig. 1).

Computational Methods—Quantification methods can be divided into feature-intensity-based methods and spectral counting methods. In the former, one tries to account for all signals corresponding to a specific charged peptide on the MS level; in the latter, one tries to infer the expression level of the peptide from the number of MS/MS identifications. Map alignment is especially important for feature-intensity-based quantification, whereas in spectral counting one can use the identification of the peptide to assign corresponding quantities across maps. Only accurate alignments of maps enable the correct comparison of quantitative properties. In the following, we describe the main steps that are necessary for label-free data processing.

Signal Processing—Depending on the type of instrument, the processing of the raw data can differ. However, there are certain generic steps in signal processing that apply to most instruments and to both intensity-based methods and spectral counting. These are *baseline filtering*, *noise filtering*, *centroiding*, and *charge estimation*.

In MALDI spectra, and to some extent in ESI spectra, a baseline is apparent that adds up to the signal caused by the analytes. In MALDI spectra, the baseline can become dominant in the low $m = z$ regions and disappears with increasing $m = z$. It is typically shaped like an exponential decay distribution and can be attributed to matrix material. The baseline leads to poorly resolved peak shapes due to a loss of baseline separation between adjacent peaks. The baseline thus interferes with intensity estimation and has to be removed computationally. Morphological filters such as the Top-hat filter can be used for this task.

In addition to the baseline signal, every mass spectrometer suffers from high-frequency noise (electronic noise, usually attributed to the detector, and chemical noise, usually attributed to solvents, buffers, and contaminants), and thus peaks expected to be approximately Gaussian in shape might not be convex any longer. This is a potential pitfall for algorithms that rely on local minima to separate isotope peaks. A noise filter will smooth the data by removing high-frequency noise—for example, a Savitzky-Golay filter will work well.

Finally, a signal that has been baseline corrected and smoothed is subjected to centroiding. The computational problem ranges here from almost trivial (*e.g.* for high-resolution spectra) to a fitting of overlapping (skewed) Gaussians, for example, in the case of highly charged ion trap signals. In general, this fitting is interwoven with the problem of obtaining the (initially) unknown charge state of a peptide, as the charge state z determines the distance of the isotope peaks, namely, $1 = z$. The resulting model fit can be used to analytically determine the peak volume and the height of the peak. Usually the peak volume is taken as the intensity of a centroided peak, because it corresponds directly to the ion count. However, for high-resolution spectra, the height of the peak (which is easier to determine) serves equally well.

Feature-based Quantification—Algorithmically, the main steps in feature-based quantification can be divided into (i) signal processing, (ii) feature finding, and (iii) map alignment. The advent of high-resolution mass spectrometers has made the signal processing and peak picking tasks simpler than they were on low-resolution instruments. However, the quantification methods are complex, and good quantification remains a challenge.

A central task in the processing of mass spectrometric data is the detection of peptide features for all ions eluted from the liquid chromatography column. Peptides elute over time from the liquid chromatography column, get ionized, and are injected into the mass spectrometer. The mass spectrometer takes new measurements in regular, small time intervals, thereby sampling the amount of the eluting ion over time, resulting in an *elution profile*. In each measurement, an ion gives rise to a typical *isotope pattern*, which is caused by its atomic composition (see Fig. 2 for examples of an elution profile and an isotope pattern). Via integration over the elution profile and isotope pattern, peptide feature intensities can be determined. In general, one can assume that the two-dimensional distribution is a product of two independent distributions. Thus, for the marginal distribution over *m/z*, similar reasoning applies as for individual spec-

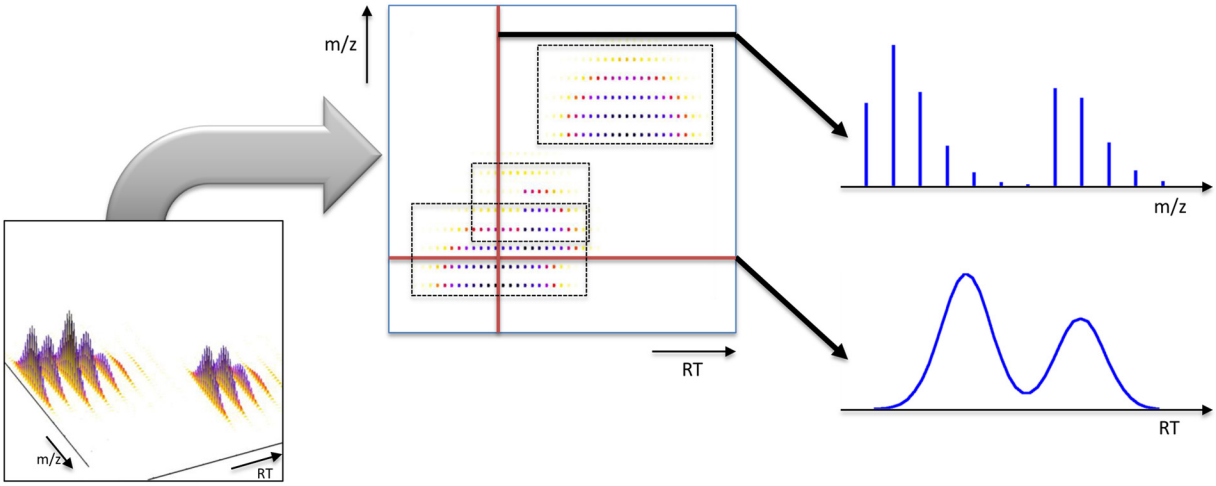


FIG. 2. **Label-free LC/MS data consist of individual MS spectra accumulated over (retention) time.** Stacked side by side, these spectra form two-dimensional maps. In these maps, individual peptides being eluted from the column give rise to sets of peaks across multiple spectra. Feature-finding algorithms can identify features, which can be defined as all mass-spectrometric signals (peaks) caused by the same peptide. Elution profiles have ideally a Gaussian shape, but they can be significantly distorted. The projection of a feature along the *m/z* axis accordingly corresponds to the isotope profile of the peptide.

tra. Automated detection of these features allows their comparison across different experiments. Fundamental to a quantitative comparison of analytes is the linear correlation of electrospray ionization intensity with ion concentration within a certain dynamic range. Most algorithms try to heuristically determine the extent and intensity of a feature by fitting appropriate distribution models to the data. This is done in areas of high signal intensity (*e.g.* by working on intensity sorted lists of peaks). The intensity of a feature can then be determined either by using the model parameters or simply by summing up all peak intensities in the feature region.

Spectral Counting—Besides this feature-intensity-based quantification method, spectral counting methods are also used for differential quantification. Despite the fact that spectral counting is commonly used to derive quantitative information at the protein level, the differential quantification of peptides builds the fundament of this concept. In the following we discuss spectral counting (SC)[1] concepts and illustrate how these concepts are involved in differential peptide quantification. SC in its simplest form counts the number of tandem spectra that are assigned to the same protein. There have been numerous studies using SC for the inference of quantitative information in label-free shotgun proteomics data. A collection of methods has recently been reviewed (37). Peptide-spectrum matchings can be used to infer differential ratios of peptides, but these methods are also gaining popularity for differential protein quantification. Methods that extend the simple SC to differential protein quantification include the protein abundance index (38); its extended version, the exponentially modified protein abundance index (39); the normalized spectral abundance factor (40); and the absolute protein expression (41). The robust intensity-based averaged ratio (RIBAR) and its extended version xRIBAR are part of a recent approach by Colaert *et al.* (36) that correlates the summed intensity of corresponding fragment spectra in two experiments and which has been shown to outperform other SC-based approaches such as the exponentially modified protein abundance index and normalized spectral abundance factor. Despite the development of novel methods to calculate protein abundance on MS/MS spectra, any approaches will struggle to reach high quantification accuracy because of the data-dependent ion sampling and dynamic exclusion list settings.

Recently, different label-free abundance measures have been compared, and their results were integrated with RNA expression data (42). Although the feature-based measure was more accurate, the authors found that, if normalized to the transcript abundance, spectral counting and feature-based methods perform equally well. Hoekman *et al.* (26) implemented a framework that allows the combination of different quantification approaches.

Map Alignment—The purpose of map alignment is to assign the same peptide features between maps for comparison. This is done using the assumption that the chromatographic elution time of a peptide, as well as its ionization behavior, stays relatively constant between measurements and that the measured *m/z* does not differ. Whereas the differences in the *m/z* are rather marginal, the shifts in the RT dimension can become very large and frequently show some nonlinearity.

Several algorithmic approaches have been used to adjust for these distortions. Lange and coauthors (43) used pose clustering techniques to find the best parameters for an affine transformation. The approach is simple and robust, but it cannot deal with nonlinear transformations. Descriptions of similar, more recent approaches can be found in Refs. 44 and 45. The approach discussed in Ref. 46 use the similarity of individual scans to compute a scan-wise alignment, whereas other methods use nonlinear functions to model the shift in retention time.

Apart from the pairwise alignment of two maps, another important aspect is grouping the correct features together across many maps. A discussion about metrics for map alignment, as well as an overview and assessment of different methods, can be found in Ref. 47.

Normalization—Once the peptide features of different maps are assigned to each other after map alignment, one needs to correct for systematic biases in the measured intensities. This is often called "intensity normalization." Normalization is a critical step in the label-free computational proteomics pipeline. It is necessary to account for variability in intensity signals (*e.g.* systematic errors in experimentation, sample preparation, chromatography, and mass spectrometry (48)). The microarray community has done extensive research in normalization procedures. In Ref. 48, Callister *et al.* compare the

[1] The abbreviations used are: MS/MS, tandem mass spectrometry; SC, spectral counting; TOPP, The OpenMS proteomics pipeline.

TABLE I
Overview of software packages for label-free quantification

Name	Platform(s)[a]	Latest version	Input format(s)	Graphical user interface	CMD	Open source[b]	Resolution[c]	Quant.	Statistical analysis
Academic/free									
MaxQuant (56)	**W**	1.2.2.5 (2011)	Thermo .RAW	+	−	No	H	MS1	
OpenMS/TOPP (34, 53)	**W**, **L**, **M**	1.9 (February 2012)	mz (ML|XML|Data)	+	+	+ (LGPL)	LH	MS1	−
pView 2 (59)	**W**, L, M	2.0 (July 2011)	mzXML, pepXML	+	−	+ (BSD)	H	MS1	+
mzMine 2 (35)	**W**, **L**, **M**	2.6 (February 2012)	mz (ML|XML|Data), ThermoRaw, NetCDF	+	−	+ (GPL 2.0)	LH	MS1	+
SuperHirn (55)	**L**, **M**	0.3 (January 2009)	mzXML, pepXML	−	+	+ (AL 2.0)	H	MS1	−
msInspect (58)	**W**, **L**, **M**	2.3 (January 2010)	mzXML, mzML (in head)	+	+	+ (AL 2.0)	LH	MS1	−
Viper (61)	**W**	3.48 (September 2011)	PEK, CSV (Decon2LS), mz (XML|Data)	+	−	+ (AL 2.0)	H	MS1	−
RIBAR/xRIBAR (36)	**W**, **L**, **M**	1.1 (May 2011)	ms_lims, .dat (Mascot)	+	−	+ (AL 2.0)	−	SC	−
Census (57)	**W**, **L**, **M**	1.72 (March 2010)	mzXML, MS1, MS2, pepXML, DTASelect	+	+	No	LH	SC, MS1	−
Corra (60)	L	3.1 (November 2010)	mzXML, pepXML	+	+	+ (AL 2.0)	LH	MS1	+
Commercial									
Mascot Distiller[d]	**W**	2.4.2 (October 2011)	mz (ML|XML), major vendors	+	+	No	LH	SC, MS1	−
SIEVE[e]	**W**	?	Thermo .RAW	+	−	No	LH	MS1	+
Progenesis LC-MS[f]	**W**	4.0 (September 2011)	mz (ML|XML), major vendors	+	−	No	LH	MS1	?
Scaffold[g]	**W**, **L**, **M**	3.3.3	Major search engines	+	+[h]	No	−	SC	+
Spectrolyzer[i]	**W**	1.0	mz (ML|XML|Data), major vendors	+	−	No	LH	MS1	+

[a] Bold and underlined text indicates the availability of binary packages; W = Windows OS, L = Linux OS, M = Mac OS.
[b] + License if applicable; AL = Apache License.
[c] Resolution: H = high, L = low (according to documentation).
[d] Matrix Science.
[e] Thermo Scientific.
[f] Nonlinear Dynamics Ltd.
[g] Proteome Software, Inc.
[h] Via ScaffoldBatch.
[i] MedicWave AB.

performance of four different normalization strategies for label-free proteomics data. They include a global normalization, linear regression, local regression, and quantile normalization. The authors found that normalization metrics need to be adapted depending on the data set. They conclude that quantile normalization has some advantages over other techniques, because no iterations are necessary and it does not force the mean to be zero (in log scale), as successive parts of the data (quantiles) are equalized from run to run. However, in their studies, linear regression models showed the best performance in most cases (49). Global normalization methods use information from all peak intensities per spectrum or run in order to scale the individual intensities. Kultima *et al.* (49) compared 10 different normalization metrics and show that linear regression that takes the analysis order into account performed best on three independent peptidomics (analysis of endogenous peptides) data sets. Wang *et al.* (50) argue that global normalization by a constant factor is feasible, but they caution that only a constant number of the most intense signals should be used for normalization if non-random missing data as a result of instrument detection limits is a concern.

Besides the publications by Kultima *et al.* (49) and Callister *et al.* (48), additional review articles discuss the issue of normalization of label-free proteomics data (51, 52).

Software packages for label-free quantification cover a wide range of normalization techniques, but each package offers only a limited set of methods. Some use normalization on individual maps (mzMine2, Corra), most use a list of matched peptide intensity pairs for normalization, and some provide no information at all. mzMine2 works on single maps and offers multiple normalization schemes (*e.g.* average intensity and maximum intensity normalization). Additionally, normalization to an internal standard that must be present in all maps is possible. Corra also operates in single raw maps and employs the LIMMA package for normalization before peak picking. MaxQuant and OpenMS' ProteinQuantifier both ensure that the median of peptide ratios is zero (in log space). pView2 uses a "median of medians" normalization. Mascot Distiller offers mean, sum, and median normalization of peptide ratios. Progenesis employs an iterative-median-of-ratios approach using a reference map. msInspect uses a linear model based on the highest intensity peptides between multiple runs. The most involved technique is implemented in Superhirn: maps are split into retention time segments, which are normalized separately. Normalization itself is performed hierarchically based on matched pairs in similar maps.

SOFTWARE PACKAGES

There is a growing collection of tools for label-free quantification implementing one or several of the techniques discussed in the preceding section. Out of the plethora of available software tools, we have selected several commercial and academic packages that are widely known and (to some extent) maintained. Table I gives an overview of computational tools, as well as information on their licenses, release dates, and input formats.

Some commercial packages such as SIEVE are restricted to the native vendor format and cannot read open community

formats like mzML, mzData, or mzXML, which can be easily converted so as to work with one other (*e.g.* via OpenMS/TOPP (34, 53) or ProteoWizard (54)). Mascot Distiller (Matrix Science), Spectrolyzer (MedicWave AB), Progenesis (Nonlinear Dynamics Ltd.), and Scaffold (Proteome Software, Inc.) support a wide range of vendor formats in addition to open formats like mzML. Most feature-based methods work on raw data and apply internal centroiding algorithms or can use centroided data directly. One exception is SuperHirn (55), which requires raw LC/MS data. All packages can deal with high-resolution data, but only some can work with low-resolution data. MaxQuant (56) and SuperHirn, for example, are specialized for high resolution, whereas OpenMS/TOPP and Census (57) can deal with both. Most tools support either SC or feature-based quantification, with Census and Mascot Distiller being the only exceptions in our lineup supporting both.

SC is supported by RIBAR/xRIBAR (36) and Census, both of which are freely available. The intrinsic details of Census are unknown, but they involve normalization for protein length and variability. Mascot Distiller and Scaffold are commercial alternatives, with the latter additionally supporting Gene Ontology term annotation. Mascot Distiller supports exponentially modified protein abundance index values, and Scaffold normalizes counts by the total count within the sample, gives access to relative and absolute counts, and allows for filtering rules.

Feature-based methods usually follow similar steps from raw data to protein expression tables (centroiding, feature finding, map alignment, and normalization, as well as protein inference) but differ in the implementation details, which are not always published, even for non-commercial tools. Progenesis and OpenMS/TOPP offer wavelet-based peak picking, suitable for low-resolution data, whereas MaxQuant fits a Gaussian curve and SuperHirn uses a simple local-maxima heuristic. Feature finding in MaxQuant is done using a graph-based approach iteratively using the best sub-graphs as predicted by an averagine model. OpenMS/TOPP uses either a wavelet approach based on an averagine model or a model-based approach on centroided data incorporating an RT shape fit and averagine models in the *m/z* dimension. For map alignment, SuperHirn uses a LOWESS fit, and OpenMS/TOPP uses a linear (affine) model or b-spline driven by either pose clustering or MS2 identification landmarks with respect to a reference map. Similarly, MsInspect (58) employs smoothing-spline regression. *Progenesis* uses a different approach of first using map alignment based on centroided data, guided by (user-defined) landmarks. Once a master map of all peak information from all maps is created, features are identified using an isotope-fitting procedure. Statistical post-processing or visualization at the protein level (where inference methods differ widely) is not supported by all tools and in this case must be diverted to dedicated statistical tools such as R. pView (59) has a tight R integration, Corra (60) features plots, and mzMine2 (35) allows for basic analysis procedures (*e.g.* PCA). Spectrolyzer has potent visualization capabilities and built-in classification and regression functionality.

Almost all packages run on Windows, with the exception of Corra and SuperHirn. Not every package provides a binary installer, so manual compilation might be required. Commercial packages tend to be Windows only; all non-commercial packages support Linux (with the exception of VIPER (61)) (see Table I for details).

CONCLUSION

Quantitative proteomics is highly relevant for systems biology, biomarker discovery, and many other biomedical applications. Among all the methods for differential peptide quantification, label-free approaches provide the highest flexibility, and as a result of recent progress in software and hardware, their dynamic range and accuracy are continuously improving. Both SC and intensity-based measures have been shown to provide good quantification results. The intensity-based measures avoid stochastic effects in ion sampling and are therefore slightly more accurate, and they potentially provide higher reproducibility. SC is easy to implement and fast.

There is a large collection of software solutions that are currently used for label-free peptide quantification, and each comes with different strengths and weaknesses. For users who intend to use standard workflows and do not need to develop algorithms and pipelines themselves, monolithic solutions such as Progenesis or MaxQuant are very suitable tools for fast data analysis. If more flexibility is needed or if an understanding of the underlying algorithms is required, open-source packages have their advantages. Large proteomics labs and core facilities will most likely appreciate the modularity and automation provided by pipeline tools.

A current challenge arises from the increasing amount of samples in more and more complex proteomics studies, in particular in clinical proteomics. Although label-free techniques scale well in general, many software tools have issues with these large-scale studies. The mere amount of data involved (hundreds of LC/MS runs resulting in hundreds of gigabytes of data) certainly causes problems, but also algorithmically there are scalability issues when these maps need to be aligned and linked. Whereas small analyses can be run on laptop computers, studies requiring more than a dozen maps usually require more powerful hardware. Multicore central processing units with a large amount of random access memory (64+ GB) and a generous amount of hard disk space are recommended for these larger studies.

Although there is still room for improvement, software tools for label-free quantification have reached a level of sophistication that makes their use convenient and reliable for most purposes. In many cases, label-free quantification

is thus a good alternative to labeling techniques in quantitative proteomics.

* O.K. acknowledges funding from the EU (FP7, PRIME-XS, and MARINA) and from BMBF (SARA, 0315395F). C.B. was supported by the European Commission's 7th Framework Program (PREDICTIV, GA202222).

¶ To whom correspondence should be addressed: Prof. Dr. Oliver Kohlbacher, Center for Bioinformatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany. Tel.: +49 7071 29 7 04 58; Fax: +49 7071 29 51 52; E-mail: oliver.kohlbacher@uni-tuebingen.de.

REFERENCES

1. Bantscheff, M., Eberhard, D., Abraham, Y., Bastuck, S., Boesche, M., Hobson, S., Mathieson, T., Perrin, J., Raida, M., Rau, C., Reader, V. r., Sweetman, G., Bauer, A., Bouwmeester, T., Hopf, C., Kruse, U., Neubauer, G., Ramsden, N., Rick, J., Kuster, B., and Drewes, G. (2007) Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nat. Biotechnol.* **25,** 1035–1044
2. Yates, J. R., Gilchrist, A., Howell, K. E., and Bergeron, J. J. M. (2005) Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell Biol.* **6,** 702–714
3. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127,** 635–648
4. Guruharsha, K. G., Rual, J.-F. o., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D. Y., Cenaj, O., McKillip, E., Shah, S., Stapleton, M., Wan, K. H., Yu, C., Parsa, B., Carlson, J. W., Chen, X., Kapadia, B., VijayRaghavan, K., Gygi, S. P., Celniker, S. E., Obar, R. A., and Artavanis-Tsakonas, S. (2011) A protein complex network of Drosophila melanogaster. *Cell* **147,** 690–703
5. Aebersold, R., and Mann, M., (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
6. Cox, J., and Mann, M., (2011) Quantitative, high-resolution proteomics for data-driven systems biology. *Annu. Rev. Biochem.* **80,** 273–299
7. Ong, S.-E., and Mann, M. (2005) Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* **1,** 252–262
8. Schulze, W. X., and Usadel, B. (2010) Quantitation in mass-spectrometry-based proteomics. *Annu. Rev. Plant Biol.* **61,** 491–516
9. Gstaiger, M., and Aebersold, R. (2009) Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat. Rev. Genetics.* **10,** 617–627
10. Munoz, J., Low, T. Y., Kok, Y. J., Chin, A., Frese, C. K., Ding, V., Choo, A., and Heck, A. J. R. (2011) The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. *Mol. Syst. Biol.* **7,** 550
11. Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology* **7,** 548
12. Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011) The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7,** 549
13. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature biotechnology* **17,** 994–999
14. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007) Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **389,** 1017–1031
15. Schmidt, A., Kellermann, J., and Lottspeich, F. (2005) A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics* **5,** 4–15
16. Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1,** 376–386
17. Krijgsveld, J., Ketting, R. F., Mahmoudi, T., Johansen, J., Artal-Sanz, M., Verrijzer, C. P., Plasterk, R. H. A., and Heck, A. J. R. (2003) Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics. *Nat. Biotechnol.* **21,** 927–931
18. Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. J. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3,** 1154–1169
19. Thompson, A., Schäfer, J. u., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G. u., Neumann, T., Johnstone, R., Mohammed, A. K. A., and Hamon, C. (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical chemistry* **75,** 1895–1904
20. Mueller, L. N., Brusniak, M. Y., Mani, D. R., and Aebersold, R. (2008) An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J. Proteome Res.* **7,** 51–61
21. Mallick, P., and Kuster, B. (2010) Proteomics: a pragmatic perspective. *Nat. Biotechnol.* **28,** 695–709
22. Daigo, K., Yamaguchi, N., Kawamura, T., Matsubara, K., Jiang, S., Ohashi, R., Sudou, Y., Kodama, T., Naito, M., Inoue, K., and Hamakubo, T. (2012) The proteomic profile of circulating pentraxin 3 (PTX3) complex in sepsis demonstrates the interaction with azurocidin 1 and other components of neutrophil extracellular traps. *Mol. Cell. Proteomics*
23. Mann, B. F., Goetz, J. A., House, M. G., Schmidt, C. M., and Novotny, M. V. (2012) Glycomic and proteomic profiling of pancreatic cyst fluids identifies hyperfucosylated lactosamines on the N-linked glycans of overexpressed glycoproteins. *Mol. Cell. Proteomics*
24. Kumar, C., and Mann, M. (2009) Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Lett.* **583,** 1703–1712
25. Käll, L., and Vitek, O. (2011) Computational mass spectrometry-based proteomics. *PLoS computational Biol.* **7,** e1002277
26. Hoekman, B., Breitling, R., Suits, F., Bischoff, R., and Horvatovich, P. (2012) msCompare: a framework for quantitative analysis of label-free LCMS data for comparative biomarker studies. *Mol. Cell. Proteomics*
27. Zhang, R., Barton, A., Brittenden, J., Huang, J. T.-J., and Crowther, D. (2010) Evaluation of computational platforms for LS-MS based label-free quantitative proteomics: a global view. *J. Proteomics Bioinform.* **3,** 260–265
28. Mayr, B. M., Kohlbacher, O., Reinert, K., Sturm, M., Gröpl, C., Lange, E., Klein, C., and Huber, C. G. (2006) Absolute myoglobin quantitation in serum by combining two-dimensional liquid chromatography-electrospray ionization mass spectrometry and novel data analysis algorithms. *J. Proteome Res.* **5,** 414–421
29. Timm, W., Scherbart, A., Bocker, S., Kohlbacher, O., and Nattkemper, T. W. (2008) Peak intensity prediction in MALDI-TOF mass spectrometry: a machine learning study to support quantitative proteomics. *BMC Bioinformatics* **9,** 443
30. Schwanhäusser, B. r., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature* **473,** 337–342
31. Krishnamurthy, D., Levin, Y., Harris, L. W., Umrania, Y., Bahn, S., and Guest, P. C. (2011) Analysis of the human pituitary proteome by data independent label-free liquid chromatography tandem mass spectrometry. *Proteomics* **11,** 495–500
32. Hyung, S.-W., Lee, M. Y., Yu, J.-H., Shin, B., Jung, H.-J., Park, J.-M., Han, W., Lee, K.-M., Moon, H.-G., Zhang, H., Aebersold, R., Hwang, D., Lee, S.-W., Yu, M.-H., and Noh, D.-Y. (2011) A serum protein profile predictive of the resistance to neoadjuvant chemotherapy in advanced breast cancers. *Mol. Cell. Proteomics* **10,** M111.011023
33. Cox, J. u., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
34. Sturm, M., Bertsch, A., Gröpl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., and Kohlbacher, O. (2008) OpenMS—an open-source software framework for mass spectrometry. *BMC Bioinformatics* **9,** 163
35. Pluskal, T. a., Castillo, S., Villar-Briones, A., and Oresic, M. (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11,** 395
36. Colaert, N., Gevaert, K., and Martens, L. (2011) RIBAR and xRIBAR: methods for reproducible relative MS/MS based label-free protein quantification. *J. Proteome Res,* 110512053707083.
37. Lundgren, D. H., Hwang, S.-I., Wu, L., and Han, D. K. (2010) Role of spectral counting in quantitative proteomics. *Expert Rev. Proteomics* **7,** 39–53
38. Rappsilber, J., Ryder, U., Lamond, A. I., and Mann, M. (2002) Large-scale

proteomic analysis of the human spliceosome. *Genome Res.* **12,** 1231–1245
39. Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **4,** 1265–1272
40. Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K., Florens, L., and Washburn, M. P. (2006) Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J. Proteome Res.* **5,** 2339–2347
41. Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E. M. (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25,** 117–124
42. Ning, K., Fermin, D., and Nesvizhskii, A. I. (2012) Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. *J. Proteome Res*.
43. Lange, E., Gröpl, C., Schulz-Trieglaff, O., Leinenbach, A., Huber, C., and Reinert, K. (2007) A geometric approach for the alignment of liquid chromatography-mass spectrometry data. *Bioinformatics* **23,** i273-i281
44. Ballardini, R., Benevento, M., Arrigoni, G., Pattini, L., and Roda, A. (2011) MassUntangler: a novel alignment tool for label-free liquid chromatographymass spectrometry proteomic data. *J. Chromatogr. A* **1218,** 8859–8868
45. Zhang, Z. (2012) Retention time alignment of LC/MS data by a divide-and-conquer algorithm. *J. Am. Soc. Mass Spectrom*.
46. Vandenbogaert, M., Thiao-Té, S., Kaltenbach, H.-M., Zhang, R., Aittokallio, T., and Schwikowski, B. (2008) Alignment of LC-MS images, with applications to biomarker discovery and protein identification. *Proteomics* **8,** 650–672
47. Lange, E., Tautenhahn, R., Neumann, S., and Gröpl, C. (2008) Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics* **9,** 375
48. Callister, S. J., Barry, R. C., Adkins, J. N., Johnson, E. T., Qian, W.-J., Webb-Robertson, B.-J. M., Smith, R. D., and Lipton, M. S. (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.* **5,** 277–286
49. Kultima, K., Nilsson, A., Scholz, B., Rossbach, U. L., Fälth, M., and Andrén, P. E. (2009) Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides. *Mol. Cell. Proteomics* **8,** 2285–2295
50. Wang, P., Tang, H., Zhang, H., Whiteaker, J., Paulovich, A. G., and Mcintosh, M. (2006) Normalization regarding non-random missing values in high-throughput mass spectrometry data. *Pac. Symp. Biocomput.* 315–326
51. America, A. H. P., and Cordewener, J. H. G. (2008) Comparative LC-MS: a landscape of peaks and valleys. *Proteomics* **8,** 731–749
52. Listgarten, J., and Emili, A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **4,** 419–434
53. Kohlbacher, O., Reinert, K., Gröpl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., and Sturm, M. (2007) TOPP—the OpenMS proteomics pipeline. *Bioinformatics (Oxford, England)* **23,** e191-e197
54. Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics (Oxford, England)* **24,** 2534–2536
55. Mueller, L. N., Rinner, O., Schmidt, A., Letarte, S., Bodenmiller, B., Brusniak, M.-Y., Vitek, O., Aebersold, R., and Muller, M. (2007) SuperHirn—a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **7,** 3470–3480
56. Cox, B., Kislinger, T., Wigle, D. A., Kannan, A., Brown, K., Okubo, T., Hogan, B., Jurisica, I., Frey, B., Rossant, J., and Emili, A. (2007) Integrated proteomic and transcriptomic profiling of mouse lung development and Nmyc target genes. *Mol. Syst. Biol.* **3,** 109
57. Park, S. K., Venable, J. D., Xu, T., and Yates, J. R. (2008) A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat. Methods* **5,** 319–322
58. Bellew, M., Coram, M., Fitzgibbon, M., Igra, M., Randolph, T., Wang, P., May, D., Eng, J., Fang, R., Lin, C., Chen, J., Goodlett, D., Whiteaker, J., Paulovich, A., and McIntosh, M. (2006) A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics (Oxford, England)* **22,** 1902–1909
59. Khan, Z., Bloom, J. S., Garcia, B. A., Singh, M., and Kruglyak, L. (2009) Protein quantification across hundreds of experimental conditions. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 15544–15548
60. Brusniak, M.-Y., Bodenmiller, B., Campbell, D., Cooke, K., Eddes, J., Garbutt, A., Lau, H., Letarte, S., Mueller, L. N., Sharma, V., Vitek, O., Zhang, N., Aebersold, R., and Watts, J. D. (2008) Corra: computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. *BMC Bioinf.* **9,** 542
61. Monroe, M. E., Tolic, N., Jaitly, N., Shaw, J. L., Adkins, J. N., and Smith, R. D. (2007) VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics (Oxford, England)* **23,** 2021–2023