ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/235421358

An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · FEBRUARY 2013

Impact Factor: 4.25 · DOI: 10.1021/pr300992u · Source: PubMed

CITATIONS
35

READS
134

11 AUTHORS, INCLUDING:

Sven Nahnsen
University of Tuebingen
13 PUBLICATIONS **228** CITATIONS
SEE PROFILE

Lars Nilse
University of Freiburg
8 PUBLICATIONS **90** CITATIONS
SEE PROFILE

Oliver Kohlbacher
University of Tuebingen
236 PUBLICATIONS **4,805** CITATIONS
SEE PROFILE

Lars Malmström
University of Zurich
60 PUBLICATIONS **1,791** CITATIONS
SEE PROFILE

Available from: Jonas Grossmann
Retrieved on: 30 January 2016

Journal of proteome research

Article

pubs.acs.org/jpr

An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics

Hendrik Weisser,[†,‡] Sven Nahnsen,[¶] Jonas Grossmann,[§] Lars Nilse,[¶] Andreas Quandt,[†] Hendrik Brauer,[¶] Marc Sturm,[¶] Erhan Kenar,[¶] Oliver Kohlbacher,[¶] Ruedi Aebersold,[†,∥] and Lars Malmström*[,†]

[†]Department of Biology, Institute of Molecular Systems Biology, ETH Zürich, 8093 Zürich, Switzerland

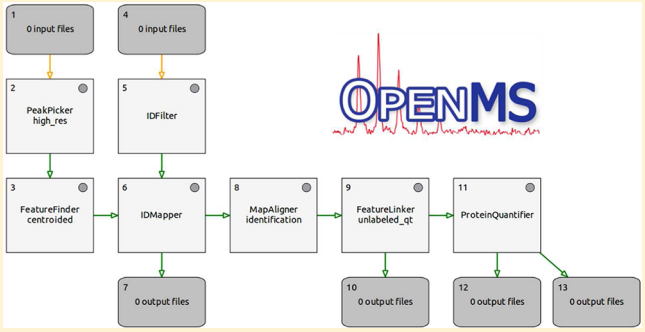
[‡]Life Science Zurich Ph.D. Program on Systems Biology of Complex Diseases, [¶]Center for Bioinformatics, Eberhard Karls University Tübingen, 72076 Tübingen, Germany

[§]Functional Genomics Center Zürich, University of Zürich and ETH Zürich, 8057 Zürich, Switzerland

[∥]Faculty of Science, University of Zürich, 8092 Zürich, Switzerland

Supporting Information

ABSTRACT: We present a computational pipeline for the quantification of peptides and proteins in label-free LC−MS/MS data sets. The pipeline is composed of tools from the OpenMS software framework and is applicable to the processing of large experiments (50+ samples). We describe several enhancements that we have introduced to OpenMS to realize the implementation of this pipeline. They include new algorithms for centroiding of raw data, for feature detection, for the alignment of multiple related measurements, and a new tool for the calculation of peptide and protein abundances. Where possible, we compare the performance of the new algorithms to that of their established counterparts in OpenMS. We validate the pipeline on the basis of two small data sets that provide ground truths for the quantification. There, we also compare our results to those of MaxQuant and Progenesis LC−MS, two popular alternatives for the analysis of label-free data. We then show how our software can be applied to a large heterogeneous data set of 58 LC−MS/MS runs.



KEYWORDS: *algorithms, automation, bioinformatics software, label-free, mass spectrometry, quantification*

INTRODUCTION

Mass spectrometry (MS) is enjoying rapid development with increased speed, sensitivity, data quality and robustness as a result. New instruments are introduced on a regular basis, new methods increase their versatility and performance, and software is becoming better, faster and easier to use. Protein identification and protein quantification are two of the most fundamental application areas in proteomics and are often carried out using so-called shotgun mass spectrometry.[1] The protein sample is enzymatically digested into peptides, typically using trypsin. The peptides are separated using one- or multidimensional liquid chromatography (LC) and are introduced into an online mass spectrometer using electrospray ionization (ESI). The mass-to-charge ratio (m/z) and intensity (in arbitrary units) is recorded for all peptide ions in a survey mass spectrum (MS1), and one or more peptide ions are selected for fragmentation using collision-induced dissociation (CID) or related methods. Resulting fragment ions are measured and recorded in a tandem MS spectrum (MS/MS or MS2). The MS2 spectra can be used to identify peptides, while the MS1 spectra allow to estimate their relative quantities.

A modern mass spectrometer can easily produce 35 000 spectra per hour of operation and is often operated around the clock, making it difficult to analyze the data manually. Rather, data analysis is typically accomplished using several specialized software applications strung together into workflows. Important types of such applications are search engines that infer peptides from MS2 spectra (e.g., Sequest,[2] Mascot,[3] OMSSA,[4] X! Tandem,[5] Inspect,[6] MyriMatch[7]), or tools that assess the statistical significance of these assignments (e.g., PeptideProphet[8] or Percolator[9]). Not least, a variety of experimental approaches and corresponding software tools can be used to derive quantitative information from LC−MS measurements. Domon and Aebersold distinguish three main quantitative proteomics strategies (shotgun/discovery, directed, and targeted proteomics) mainly based on the use of prior information in the quantification workflow.[10] Different labeling schemes (reviewed, e.g., by Bantscheff et al.[11]) provide a second and orthogonal criterion for categorization.

Here, we discuss label-free quantification, a technique for quantifying peptides and proteins in unlabeled samples. We specifically apply the term for quantification methods that are based on integrating ion currents extracted from MS1 spectra of

Received: October 22, 2012
Published: February 8, 2013

ACS Publications © 2013 American Chemical Society

dx.doi.org/10.1021/pr300992u | J. Proteome Res. 2013, 12, 1628−1644

shotgun measurements, in contrast to spectral counting approaches[12] or to the quantification of unlabeled samples via selected reaction monitoring (SRM).[13,14] In practice, label-free quantification (in this stricter sense) works by detecting peptide features in the LC−MS data, integrating the signal intensities within each feature, and if possible, attributing this intensity to a peptide identified from an MS2 spectrum. In many cases, there is also an alignment step that merges data from related samples of one experiment, so that information can be inferred across different measurements and samples. In the end, feature intensities are accumulated to peptide and protein abundances.

Several properties make the label-free approach an attractive choice for quantitative proteomics: Costs and efforts that would be associated with the labeling of samples are saved, and there is no artificial increase in sample complexity due to the mixing of differently labeled samples. It is not necessary to define target peptides or proteins a priori, as is the case with SRM, where specific assays need to be developed and validated for each peptide. Indeed, label-free quantification is well suited for the task of identifying targets of interest for subsequent analyses, e.g., in the discovery phase of biomarker studies.[15] The minimum of required preparation, combined with the fact that many hundreds to several thousands of proteins can be quantified per LC−MS measurement, make label-free quantification ideal for high-throughput studies that aim to rapidly quantify the contents of large numbers of samples.

There are also downsides to label-free quantification of shotgun data: (i) The lack of an internal standard means that data needs to be normalized carefully to become comparable across different measurements. However, this can be overcome by the addition of stable isotope-labeled standard peptides to the samples, which even allows for absolute quantification.[16] (ii) Data sets collected over long periods of time are also difficult to analyze together since they might have large retention time shifts. (iii) The limit of detection is not as low and the dynamic range is not as high as with SRM, resulting in low-abundance peptides/proteins being quantified less reliably. Moreover, since feature detection is not perfect, the fact that a peptide could not be detected in a sample does not necessarily imply that its abundance is below the detection level. Consequently, large data sets collected from complex samples such as cell lysates are often associated with missing data in the form of proteins or peptides that are detected and quantified only in a subset of the samples. This makes subsequent data analysis more difficult. Thus, there is still room for improvement despite the numerous successful studies published in the past few years.

Popular software tools for label-free quantification include SuperHirn,[17] MaxQuant,[18] Progenesis LC−MS (by Nonlinear Dynamics), and OpenMS/TOPP.[19,20] OpenMS is an open source effort with a large user base developed in collaboration between several universities. The OpenMS framework provides applications for a wide variety of data processing tasks, the so-called TOPP (short for "The OpenMS Proteomics Pipeline") tools, that can be chained into complex workflows. Each TOPP tool is dedicated to a specific task, but some tools offer a choice of different algorithms, e.g., for different types of input data or using different computational approaches. The core steps of the label-free pipeline are implemented in the TOPP tools FeatureFinder (feature detection including integration of signal intensities), IDMapper (combination of features and identified peptides), and MapAligner/FeatureLinker (alignment). The PeakPicker tool may be used as a preprocessing step to transform raw data from profile mode to centroid mode (centroiding).

In this paper, we present a software pipeline for accurate label-free quantification, which is suitable for the automated processing of large-scale data sets. The pipeline is implemented using TOPP tools from OpenMS 1.8, and we describe the numerous improvements compared to the corresponding tools in earlier versions of OpenMS. We introduce new algorithms that we developed for the TOPP tools that perform centroiding (PeakPicker), feature detection (FeatureFinder), retention time adjustment (MapAligner), and feature grouping (FeatureLinker), as well as a new TOPP tool for quantification on the level of peptides and proteins (ProteinQuantifier). We demonstrate the enhancements by analyzing two smaller data sets that provide ground truths for the quantification, as well as a large data set that focuses on the high-throughput aspect. We compare our OpenMS pipeline to MaxQuant and Progenesis, and where applicable compare new algorithms to their established alternatives in OpenMS.

■ EXPERIMENTAL PROCEDURES

Data Sets

To validate our label-free data analysis pipeline, we applied it to two data sets that allow us to compare the quantitative results to a ground truth: The first data set, termed "dilution series", consists of six samples in which whole cell lysates of the bacterium *Streptococcus pyogenes* and of human cells were mixed in different ratios. To prepare the samples, six independent biological samples of the *S. pyogenes* strain SF370 were grown to log phase (three samples) and to stationary phase (three samples), then harvested, digested, and pooled. In addition, one confluent sample of the human cell line HFL-1 (human fetal lung fibroblasts) was grown, harvested, and digested. The pooled *S. pyogenes* samples and the HFL-1 sample were each measured on an LTQ-Orbitrap XL instrument; the total ion currents (TICs) were recorded. On the basis of the recorded TICs, the bacterial and the human sample were mixed in a dilution series with the ratios 0/100, 20/80, 40/60, 60/40, 80/20, and 100/0%. The six samples of the dilution series were measured a single time each on an LTQ-Orbitrap instrument (75 min gradient), using data-dependent acquisition (DDA) to select precursors for MS2 fragmentation. MS1 scans were acquired in profile mode. For more details on data acquisition, see Malmström et al.[21] The data was searched with Mascot, OMSSA, and X! Tandem against a combined human−*S. pyogenes* target/decoy protein database. After performing label-free quantification as described below, we normalized the results by scaling the peptide abundances to equal sums in every sample, on the basis of the premise that the same total amount of protein should be present in each sample.

The second data set, termed "Leptospira", consists of four samples from the bacterium *Leptospira interrogans*. This is a part of the data prepared and analyzed by Schmidt et al.[16] (see there for further details). Two untreated control samples, one sample of treatment with the antibiotic ciprofloxacin for 12 h, and one sample of ciprofloxacin treatment for 24 h were measured in technical duplicates on an LTQ-FT instrument (140 min gradient). MS1 scans were acquired in profile mode. The technical replicates used different inclusion lists to control the selection of precursors for MS2. The data was searched with OMSSA, X! Tandem, and MyriMatch against a target/decoy database of the predicted *L. interrogans* proteome. In addition,

dx.doi.org/10.1021/pr300992u | J. Proteome Res. 2013, 12, 1628−1644

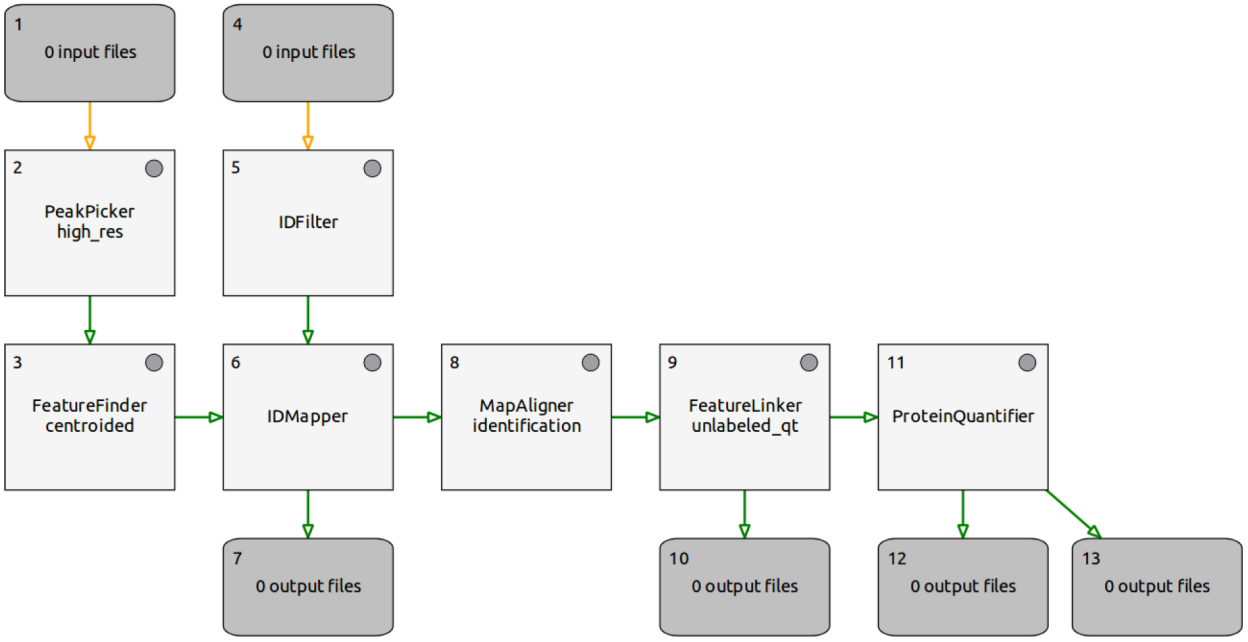


Figure 1. Core OpenMS label-free analysis pipeline, as it appears in TOPPAS—OpenMS' GUI for the design of TOPP workflows. The inputs and outputs are as follows. 1: mzML files containing LC—MS/MS data in profile mode. 4: idXML files containing protein database search results. 7: featureXML files containing feature maps annotated with peptide IDs. 10: consensusXML file containing the consensus map produced from all input files. 12/13: CSV files containing peptide and protein abundance data.

in the same samples 39 proteins covering a wide abundance range were quantified using SRM by Ludwig et al.[14] We use these SRM results as a gold standard for quantification. After the label-free data analysis, we again normalized the results; here, abundances were scaled so that the median peptide abundance was the same in every sample.

Furthermore, in order to assess the feasibility of analyzing large amounts of data in a high-throughput fashion, we tested a data set of 58 measurements from three related experiments. This data set is termed "Streptococcus", because it captures *Streptococcus pyogenes* grown under different conditions. The data comes from a study by Malmström et al.,[21] which includes detailed descriptions of the samples and the data acquisition. In summary, three experiments were performed: In the "plasma proteins" experiment, two independent biological samples were prepared for each of five different growth conditions. Each sample was measured in three technical replicates, using DDA and two distinct inclusion lists. In total, 30 measurements were acquired for this experiment. In the "dose response" experiment, there were again two biological replicates for each of five growth conditions, but in this case only one DDA measurement per sample was performed. This experiment consists of 10 measurements. Finally, the "fatty acids" experiment includes six growth conditions and three biological replicates per condition. Each sample was measured once in DDA mode, for a total of 18 LC—MS measurements. All measurements were performed on the same LTQ-FT instrument, with a 110 min gradient for the "plasma proteins" and "fatty acids" experiments, and a 140 min gradient for the "dose response" experiment. In all cases, MS1 scans were acquired in profile mode. We searched the data with OMSSA and X! Tandem against a target/decoy database of *S. pyogenes* M1 GAS proteins (RefSeq genome annotation retrieved from PATRIC[22]). There is no ground truth for this data set.

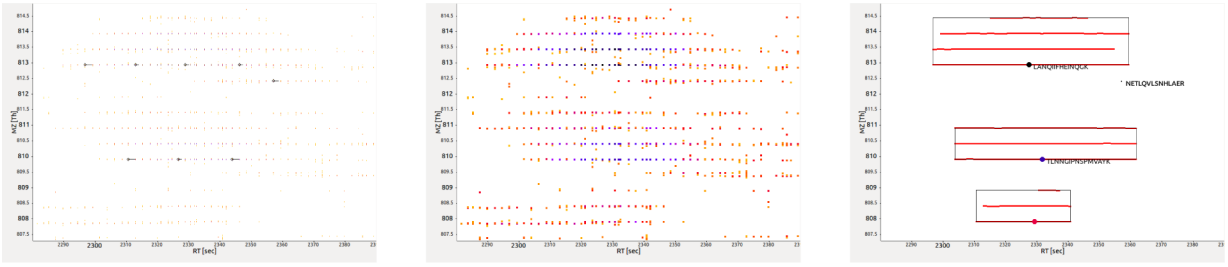
Processing of Peptide Identifications

Label-free quantification methods derive quantitative information from integrated signal intensities in MS1 spectra. In order to attribute the quantitative information to peptides and proteins, peptide identifications based on MS2 spectra (peptide-spectrum matches) are needed. We will not discuss the problem of identifying MS2 spectra here, since it is separate from the problem of quantification. In practice, peptide identifications typically come from protein database search engines (e.g., Inspect,[6] Mascot,[3] OMSSA,[4] X! Tandem[5]) or spectral library search engines (e.g., SpectraST[23]).[24] OpenMS provides a number of adapters to interface with different search engines, as well as a converter from the pepXML file format used by the popular Trans-Proteomic Pipeline (TPP)[25] to OpenMS' own idXML format.

For most of the well-established workflows, measuring a reasonably complex sample on a hybrid mass spectrometer in DDA mode, the number of detected features is usually much higher than the number of identified spectra because of limited scan rates. In order to quantify the contents of a complex sample on the peptide or protein level as completely as possible, it is thus critical to maximize the number of identified MS2 spectra without sacrificing accuracy. Against this background, approaches that combine results from different protein identification engines to achieve higher identification rates and to improve confidence in the matches become especially valuable.

Our pipeline for label-free quantification can be interfaced with sources for peptide identifications that integrate results from multiple search engines. One example is the iProphet program[26] that is part of the TPP. To use combined search results from a TPP measurement in OpenMS, the pepXML file produced by iProphet can be read and converted to idXML with the TOPP tool IDFileConverter. Currently, OpenMS does not include functionality for statistical validation of identi-

dx.doi.org/10.1021/pr300992u | J. Proteome Res. 2013, 12, 1628—1644



(A) Profile LC-MS map. (B) Centroided LC-MS map. (C) Annotated feature map.

Figure 2. Data processing stages illustrated on a section of the first run of the "Leptospira" data set (images exported from TOPPView). (A) Profile LC−MS map. Colors indicate local intensity, MS2 precursors are marked by black diamonds and lines. (B) Centroided LC−MS map, after processing with PeakPicker ("high_res" algorithm). (C) Annotated feature map, after processing with FeatureFinder ("centroided" algorithm) and IDMapper. Depicted are two features annotated with peptide identifications, one unannotated feature, and one unassigned identification. Extracted mass traces of the features are shown in red.

fication results on the protein level, like ProteinProphet[27] in the TPP. However, ProteinProphet results in protXML format can also be converted by IDFileConverter and can subsequently be taken into account for protein quantification.

Alternatively, a TOPP workflow constructed around the ConsensusID tool[28] can be used to rescore and combine results from multiple search engines. An example workflow is shown in Figure S1 (Supporting Information). Independent of their source, peptide identifications used for label-free quantification should be filtered to a high-confidence subset, e.g., using a false discovery rate (FDR) cutoff of 1%. This is possible with the TOPP tool IDFilter.

OpenMS Data Processing

We processed the test data sets with the label-free pipeline outlined in detail below, implemented in OpenMS 1.8. As file format for input and output of LC−MS/MS raw data, OpenMS uses mzML, the designated standard format developed by the HUPO Proteomics Standards Initiative.[29] Conversion of a number of open file formats (e.g., mzXML) to/from mzML is supported by the TOPP tool FileConverter. Conversion of proprietary binary file formats (e.g., Thermo RAW) to mzML is not possible with OpenMS 1.8 but should be supported by software from the respective vendor. As our source of peptide and protein identifications, we used TPP results (via conversion of pepXML/protXML files to idXML format with IDFileConverter). With the IDFilter TOPP tool, we filtered the peptide IDs based on iProphet scores to a high-confidence subset of 1% FDR.

The "dilution series" data set was analyzed with the "basic" quantification pipeline as depicted in Figure 1. For the "Leptospira" data set, we performed an additional preprocessing step: Since every sample was measured twice, with two different inclusion lists, the sets of peptide IDs associated with the two technical replicates of a sample were largely (but not completely) distinct. We combined the two corresponding sets of IDs by aligning (MapAligner "identification") and merging (IDMerger) them, so that the full set of IDs from both inclusion lists could be used to annotate each replicate measurement of a sample. For the "Streptococcus" data set, we used a hierarchical approach: We first analyzed the three subexperiments individually with our "basic" pipeline. For each experiment, we obtained a so-called consensus map as outcome of the FeatureLinker step. We then combined all results by performing an alignment (MapAligner followed by FeatureLinker) of the three consensus maps. In the RT adjustment step, we used the consensus map from the "dose response" experiment as reference, because of its longer gradient. In the end, we performed peptide and protein quantification based on the merged results from all experiments.

The parameter settings of TOPP tools that we changed from their default values in OpenMS 1.8 to process the three data sets are listed in Table S1 (Supporting Information). We chose settings that result in a very sensitive detection of features (PeakPicker and FeatureFinder parameters), even more so for the "Leptospira" and "Streptococcus" data sets that were measured on an LTQ-FT instrument using inclusion lists. However, in later steps we applied quite stringent parameters to avoid false positive matches (IDMapper and FeatureLinker). For the "Streptococcus" data set, we adapted the parameters of the alignment tools (MapAligner and FeatureLinker) to account for the higher number of runs to be aligned. For quantification on the protein level (ProteinQuantifier), we used the single proteotypic peptide of highest abundance.

Comparison to Alternative Software

In addition to our OpenMS pipeline, we processed the example data sets with two alternative software solutions for label-free quantification, the freely available MaxQuant[18] and the commercial Progenesis LC−MS (by Nonlinear Dynamics). Both tools require the operating system Microsoft Windows, in contrast to OpenMS, which supports Linux, Windows, and Mac OS. Both MaxQuant and Progenesis operate on LC−MS raw data in profile mode, which is read directly from Thermo RAW files (although Progenesis also supports the open format mzXML).

For the MaxQuant analysis, we used MaxQuant version 1.1.1.25. The software includes its own search engine Andromeda,[30] which was used for peptide and protein identification in the MaxQuant case. Search results were always filtered to 1% FDR. We configured MaxQuant for label-free quantification (including matching between runs), and provided "experimental design templates" indicating the replicate inclusion list measurements in the "Leptospira" and "Streptococcus" data sets. We extracted peptide and protein quantification results from the "peptides.txt" and "proteinGroups.txt" files produced by MaxQuant for comparison with the OpenMS and Progenesis results.

We further analyzed our example data sets with Progenesis LC−MS (version 4.0 and above) as follows: We chose Progenesis machine settings according to the respective acquisition instrument. In the alignment step, we manually

dx.doi.org/10.1021/pr300992u | J. Proteome Res. 2013, 12, 1628−1644

selected a reference run and supplied three to five seed vectors per other input run, spread over the whole retention time range, to serve as a good starting point for the automatic alignment. We selected the most sensitive setting ("5") for the feature detection algorithm in all cases. As source of peptide and protein identifications, we imported the TPP results that we also used in the OpenMS pipeline (filtered to 1% FDR). We exported the quantification results either on the protein level or on the feature level. In the latter case, we aggregated feature data to peptide data in postprocessing by summing up all normalized intensities of features annotated with the same peptide sequence.

RESULTS

The OpenMS Label-Free Quantification Pipeline

Below, we describe the different parts of our label-free quantification pipeline and the TOPP tools and algorithms used to implement them. Figure 1 gives an overview of the core of the pipeline as it appears in TOPPAS, the "TOPP Pipeline Assistant" of OpenMS, which provides a graphical user interface for the creation and execution of TOPP workflows.[31] TOPPAS makes it easy to design even complex workflows and allows data processing at the press of a button. While our specific implementation of a label-free quantification pipeline is one of several variants that could be constructed in OpenMS for the task, we found it well suited for most common use cases. Inputs of the pipeline are LC−MS/MS raw data (MS1 in profile mode) in mzML format and peptide/protein identifications in idXML format. The goal is to accurately quantify as many as possible of the peptides and, by extension, proteins that have been identified with high confidence. Figure 2 illustrates principal stages of the data processing on an example.

PeakPicker ("high_res"): Fast Centroiding of High-Resolution Data. Centroiding, also called "peak picking", is a preprocessing step applied to continuous ("profile") mass spectra. To reduce the amount of data and to simplify further processing, the peaks in a continuous spectrum are detected, and every peak is represented by a single data point (of appropriate intensity) located at its apex, the centroid, which does not need to coincide with a sampling point of the original spectrum. While it is often possible to perform centroiding with software provided by the vendor of the mass spectrometer, using the TOPP tool PeakPicker ensures adequate results for subsequent processing with OpenMS.

For high-resolution data, such as that produced by FT-ICR or Orbitrap instruments, we have developed the "high_res" PeakPicker algorithm. For lower-resolution data, such as that from time-of-flight (TOF) instruments, OpenMS includes the more generally applicable "wavelet" algorithm.[32] Compared to the "wavelet" algorithm, the "high_res" algorithm is no less accurate, but faster (by a factor of 6−10), and it does not distort the intensity scale of the data. These advantages are possible because peaks in high-resolution data are typically well-defined with narrow shapes and can be clearly separated, enabling a less meticulous centroiding method. Briefly, the "high_res" algorithm applied to a spectrum works as follows: In the first phase, a region of neighboring data points that make up a significant peak is identified and extended as far as possible. Such a region must contain at least three data points and satisfy the following conditions: (i) All points pass a signal-to-noise filter (with a user-defined threshold); (ii) the points are approximately evenly spaced in m/z; and (iii) their intensities are first strictly increasing, up to a maximum, and then strictly decreasing. In the second phase, an interpolating cubic spline is fitted to the points of the peak. In the third and final phase, the bisection method is applied to find a root of the spline's derivative. The root gives the m/z value of the peak's apex, and thus the position of the centroid representing the peak; the value of the spline at this position yields the corresponding intensity value. The centroid is stored, and the algorithm continues to find the next peak region, etc.

In the case of data from a hybrid mass spectrometer, where MS2 spectra were acquired on the low-resolution mass analyzer, it is possible to selectively apply the "high_res" PeakPicker only to the MS1 spectra. Since we use centroiding as a preprocessing step for subsequent feature detection on MS1 level, there is no need to centroid MS2 spectra.

FeatureFinder ("centroided"): Feature Detection on Centroided Data. Feature detection is at the heart of label-free data analysis. A (peptide) feature is a region in the LC−MS map where signals are generated by one peptide species in a specific charge state. Integrating the signal intensities in the region gives a quantitative value for the feature, which correlates with the abundance of the peptide ions. Features are characterized by an isotope pattern in the m/z dimension and an elution profile in the RT dimension. The aptly named TOPP tool "FeatureFinder" exploits these properties for the detection of features. Here, we introduce an algorithm for centroided data, termed "centroided", which is used in our pipeline. The algorithm works in five phases: (1) seeding, (2) extension, (3) model fitting, (4) feature creation, and (5) conflict resolution. We explain the phases briefly below; for more details, see Sturm.[33]

Phase 1: Seeding. The seeding phase attempts to find peaks in the LC−MS map that are likely to be extendable to features later on. It looks for peaks that are of significant intensity, are the local intensity maximum of a mass trace (several peaks of the same m/z in consecutive spectra), and are part of a peptide-like isotope pattern. To this end, scores that capture these three properties are calculated and stored for every peak in the LC−MS map: The first is the intensity score that compares the peak's intensity to the distribution of intensities in its surroundings. The second is the mass trace score that evaluates whether peaks of the same m/z occur in the preceding and following spectra. Here it is also marked whether a peak is a local intensity maximum within its mass trace. For the third score, the isotope pattern score, the averagine model[34] is used to precalculate theoretical isotope distributions of peptides of varying masses. The following steps (up to the "conflict resolution" phase) have to be carried out separately for each charge state that is considered for peptides in a specific data set. The charge-specific isotope pattern score is computed for every peak by comparing the theoretical isotope pattern (of the appropriate peptide mass and adjusted to the current charge state) to the other peaks surrounding the query peak in its spectrum and, to increase robustness, in the two adjacent spectra. The final seed score for a peak is computed as the geometric mean of intensity score, mass trace score, and isotope pattern score. All peaks from the LC−MS map that pass a seed score cutoff and constitute local intensity maxima are stored as seeds. The phases of extension, model fitting, and feature creation are carried out for each seed individually, in order of decreasing seed scores.

Phase 2: Extension. The goal of the extension phase is to collect all the peaks around a seed that potentially belong to a

dx.doi.org/10.1021/pr300992u | J. Proteome Res. 2013, 12, 1628−1644

feature. The seed in question is first extended in m/z dimension: The appropriate theoretical isotope pattern is again compared to the region around the seed, to find the best isotope pattern containing the seed and to determine how many isotopes were detected there. Next, this isotope pattern is extended in RT dimension. Starting with the isotope of highest intensity, the mass traces are extended in both directions (increasing and decreasing RT) by collecting peaks of matching m/z until (a) there are no more such peaks, (b) the noise level is reached, or (c) intensity values start increasing again, suggesting the beginning of a new feature. Mass traces of subsequent isotopes are also not allowed to extend further than the highest-intensity isotope.

Phase 3: Model Fitting. In this phase, a model for the elution profile is fitted to the peaks collected in the extension phase. Users can choose between a Gaussian model (for symmetric elution profiles) or an exponential-Gaussian hybrid model (for asymmetric elution profiles).[35] One model is fitted to all mass traces of a potential feature simultaneously. To this end, the traces are scaled according to the relative isotope intensities of the best matching theoretical isotope pattern determined in the extension phase.

Phase 4: Feature Creation. On the basis of the model fit, the feature region is trimmed in the RT dimension, and poorly matching mass traces are removed. Both the model and the resulting feature candidate are then checked for validity using several constraints. Comparing the intensity profile of the feature candidate to the theoretical profile given by the model yields a feature quality score that measures the goodness of fit. The score takes into account the relative deviations in intensities as well as the correlation of the intensity profiles. If all validity checks pass and if the feature quality exceeds a threshold, a complete feature is created. This feature is characterized by its monoisotopic m/z and charge (both derived from the isotope pattern), its retention time apex (read from the model of the elution profile), its intensity (computed as the area under the curve of the model), and its quality (the feature quality score). Furthermore, hulls for the individual mass traces are computed and added to the feature, which is then stored. Seeds within the mass trace hulls of the feature are removed from consideration, and the cycle continues with the next seed that is extended and potentially turned into a feature.

Phase 5: Conflict Resolution. Since features from different seeds are created independently of each other, some peaks in the LC−MS map may be incorporated in multiple features. The final phase of the feature detection resolves such conflicts. Features with overlapping mass traces are identified, and if the relative overlap of two features exceeds a user-defined threshold, one of them is removed. The decision which feature to remove is governed by rules involving the charge states, quality scores, and intensities of the respective features.

Through the combination of the described steps, a typical LC−MS map comprising millions of individual peaks of varying intensities is turned into a consistent set (of potentially several thousands) of discrete features. Those features correspond to distinct peptides in different charge states and capture relevant quantitative information.

Targeted Feature Detection with User-Defined Seeds. The "centroided" FeatureFinder supports user-defined seed lists, which allow to target the feature detection to specific regions of interest in the LC−MS map. In essence, a custom seed list contains positions (denoted as pairs of RT and m/z values) in the LC−MS map where features are expected. Since the seeds used internally by the "centroided" algorithm have to fulfill several criteria (see "Phase 1: Seeding" above), user-supplied seeds cannot directly replace the internally computed ones. Instead, the internal list of seeds is filtered based on proximity to user-defined seeds, using RT and m/z tolerances that can be set by the user. Thus, targeted feature detection will not find additional features that could not have been detected otherwise, but will largely limit the result to a more relevant subset of features. Some exceptions may occur where features that would have been removed during the "conflict resolution" phase of a standard feature detection are still present in a seed list-derived result. In a typical application, the use of a seed list can dramatically reduce the runtime of the feature detection and of subsequent steps, and/or allow the use of less strict parameters in order to increase sensitivity (see "Feature Detection Performance" below for an example).

Running a targeted feature detection is as simple as supplying the seed list to FeatureFinder via the seeds option. OpenMS uses the featureXML file format to store seed lists, but a tab-separated text file containing seed positions can be converted to featureXML with the FileConverter tool (conversion in the other direction is possible with the TextExporter). Seed lists can also be automatically generated by the TOPP tool SeedListGenerator on the basis of different kinds of input data. For example, given an idXML file, a seed list containing the locations of peptide identifications is created. This list can be used to focus the detection on features that can later be annotated with identifications.

IDMapper: Annotation of Features with Identified Peptides. To quantify peptides and proteins, quantitative information (features) and identification information (peptides identified from MS2 spectra) need to be integrated. In OpenMS this is done by the IDMapper TOPP tool, which annotates features with peptide identifications based on matching positions in RT and m/z (taking into account user-specified tolerances) and matching charge states. While the function of this tool has not changed fundamentally since its addition to OpenMS, we have, for the present project, optimized its runtime and improved its usability by reworking the parameter handling. For example, it is now straightforward to specify that the m/z value of a peptide should be computed from its theoretical mass, not taken from the precursor m/z of the corresponding MS2 spectrum; that this m/z value is to be compared against the monoisotopic m/z of a feature, not the whole m/z range of the feature; and that in the RT dimension, the whole range of a feature is to be considered, not just its apex retention time. These are the settings we use in our pipeline.

MapAligner ("identification"): Identification-Based Retention Time Adjustment. Working with LC−MS data from multiple measurements of related samples, it is often necessary to adjust the retention time scales of the runs to one common scale, in order to remedy limited reproducibility of the chromatography. This so-called alignment step may, for example, be necessary before related features from different runs can be grouped together, or before a run can be annotated with peptide identifications obtained in a different run. OpenMS contains the TOPP tool MapAligner for this task. When feature data from different runs should be combined, as in the pipeline presented here, the MapAligner works in connection with the FeatureLinker, described below. The alignment step then serves to reduce retention time differences

dx.doi.org/10.1021/pr300992u | J. Proteome Res. 2013, 12, 1628−1644

between runs to within the tolerance of the feature grouping step.

We describe here a new alignment algorithm, termed "identification", which was developed for the present pipeline. It is a useful complement to the alternative in OpenMS, the "pose_clustering" algorithm developed by Lange et al.,[36] because it relies on identification data instead of feature data. It is also faster and has significantly fewer parameters. The "identification" algorithm uses peptide sequences identified in different runs to establish points of correspondence between the runs (see also Fischer et al.[37] for an implementation of this idea). The median retention time (RT) of each peptide species in a run is contrasted with a reference RT, either the median RT of the same peptide in a user-defined reference run, or the median RT over all runs (like in the XCMS software for metabolomics LC−MS data analysis[38]). From these RT pairs, a function is estimated that transforms the RT scale of the individual run to the reference time scale, either the time scale of the reference run, or an average time scale of all runs. To protect against outliers, it is possible to restrict the set of peptides used in the alignment by setting a maximum allowed RT shift or a minimum number of runs in which peptides must occur.

In OpenMS 1.8, users can choose between different models for the computation of the transformation from the RT pairs. In our pipeline, we use a smooth nonlinear transformation that is calculated by fitting a cubic B-spline to the data. However, if it is not necessary to account for nonlinear differences between the retention time scales, the more robust linear model may be chosen instead.

FeatureLinker ("unlabeled_qt"): Feature Grouping by QT Clustering. The feature linking step combines feature maps from multiple related LC−MS runs to one so-called consensus map (also called "master map", e.g., in SuperHirn). This is done by finding corresponding features in different input maps and grouping them into consensus features. A consensus feature consists of a centroid, which is a derived feature representative of the whole group, and of the grouped subfeatures. Features in a group are assumed to be caused by the same peptide species, and are characterized by matching retention times, mass-to-charge ratios, charge states, and peptide annotations (if any). OpenMS stipulates that a consensus feature should only contain one subfeature from each constituent feature map; while this is not optimal for dealing with cases of features being inadvertently split up during feature detection, it considerably simplifies the feature grouping process.

In addition to the feature linking algorithm for unlabeled data described previously,[36] termed simply "unlabeled", we introduce here a new algorithm based on a variant of QT clustering,[39] termed "unlabeled_qt". Where the "unlabeled" algorithm builds up the consensus map step by step by pairwise merging of feature maps, the "unlabeled_qt" algorithm considers all inputs at once and is thus able to produce a grouping solution that is closer to the global optimum.

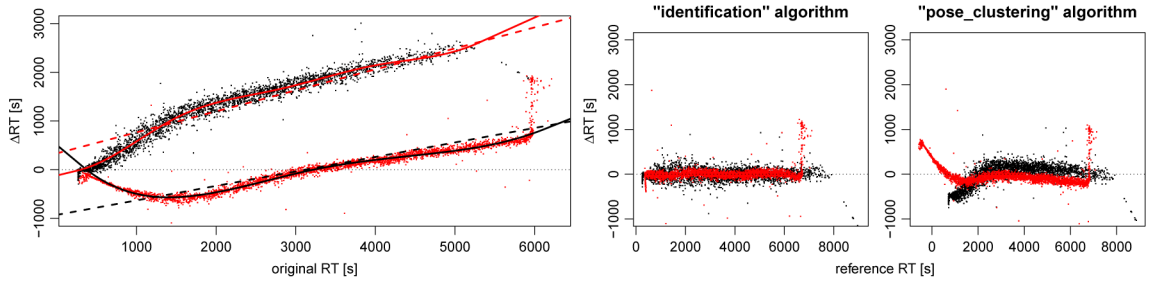
The "unlabeled_qt" algorithm works as follows: We consider every feature from every input map as a potential cluster center. Neighboring features from other maps are tentatively included in a cluster as long as they meet a quality threshold (hence the name "QT clustering"). In our case, the quality threshold is given by user-defined tolerances for the differences in RT and m/z between a feature and the cluster center, and by the further requirements of matching charge states and peptide identifications (if applicable). In contrast to the original QT clustering algorithm, the eligibility of a feature for a cluster only depends on that feature and on the cluster center. This difference allows us to perform only one full round of clustering, storing all eligible features for every cluster. For each cluster, we then find the cluster elements (features) that would produce the best consensus feature. In the simplest case, we select the feature from each map that is closest to the cluster center according to a user-adjustable distance function that incorporates differences in RT, m/z, and intensity. We can then compute the quality of the cluster based on a contribution for every input map, either the distance of the selected feature to the cluster center, or a maximum distance if there is no feature from the respective map. However, if peptide identifications have to be taken into account, we must observe that only compatible features (the same ID or no ID) may occur together in a consensus feature. If a feature has multiple different IDs mapped to it, it may only be grouped with other features with the same combination of IDs (however, these cases of ambiguous annotations are rare in practice). If the ID for the whole cluster is not already determined by the center feature, we may have to consider several possibilities. In such a case, we choose the peptide identification and respective cluster elements that lead to the highest cluster quality. We subsequently extract the cluster with the highest quality score, which is the most compact of the clusters with the highest number of compatible features. We turn that cluster into a consensus feature, comprised of the cluster center and the previously selected cluster elements. Next, we update the remaining clusters to account for the features that were removed, and recompute cluster qualities where necessary. We repeat the process of extracting the best cluster, generating a consensus feature from it, and updating the remaining clusters until no more clusters are left. The need for only one full round of clustering, together with a two-dimensional hashing of features and the usage of a look-up table for distances between features, allows us to run this extensive feature grouping approach efficiently.

ProteinQuantifier: Calculation of Peptide and Protein Abundances. The pipeline described so far allows the quantification of peptide features; however, for biological interpretation, it is often important to quantify on the level of peptides and proteins. In OpenMS 1.7 we introduced the ProteinQuantifier tool for this purpose. Given a feature map or a consensus map as input, ProteinQuantifier produces tables with peptide and protein abundances in the CSV format, which can be easily imported into, e.g., Microsoft Excel or the statistical computing environment R.

Quantification is based on the intensity values of the features in the input. Feature intensities are first accumulated to peptide abundances, according to the peptide identifications annotated to the features (or feature groups, in the case of a consensus map). Then, abundance values of the peptides of a protein are averaged to compute the protein abundance. In this peptide-to-protein step, a fixed number (defined by the user) of the most abundant proteotypic peptides, i.e., peptides matching to exactly one protein, per protein are considered for the abundance calculation. This is a general version of the "high-flier approach" described previously,[40,41] which uses the three most intense peptides per protein.

Only features/feature groups with unambiguous peptide annotation are used for peptide quantification, and generally only proteotypic peptides are used for protein quantification.

dx.doi.org/10.1021/pr300992u | J. Proteome Res. 2013, 12, 1628−1644



(A) Before the alignment. (B) After the alignment.

Figure 3. Alignment (MapAligner) of consensus maps from the three subexperiments of the "Streptococcus" data set, using the "dose response" experiment as reference. Data points correspond to peptide sequences that occur in both the reference and in the "fatty acids" (black dots) or the "plasma proteins" (red dots) experiment. For each peptide, we plot the difference to the reference retention time against the median retention time in the respective consensus map. (A) Before the alignment. Continuous red/black lines show the fit of the B-spline transformation estimated by the "identification" algorithm; dashed red/black lines show the fit of the affine transformation estimated by the "pose_clustering" algorithm (colors were reversed for better visibility). (B) After alignment with the "identification" (left) and with the "pose_clustering" (right) algorithm. (The vertical "line" of red dots near the end of the retention time scale corresponds to a set of peptides that elute together at the end of the gradient in the "plasma proteins" experiment but that were separated by the longer gradient of the "dose response" experiment. It is not possible to resolve such cases with alignment approaches like ours that "dewarp" retention time scales.).

As an exception to this rule, if ProteinProphet results for the whole sample set are provided as an additional input to ProteinQuantifier, or if they are already included in a feature map, also groups of indistinguishable proteins can be quantified. The reported quantity then refers to the total for the whole group.

Peptides with the same sequence but with different modifications are quantified separately on the peptide level but treated as one peptide for the protein quantification, i.e., the contributions of differently modified variants of the same peptide are accumulated.

Comparison of Identification Pipelines

We compared the identification results obtained for MS2 spectra from the "dilution series" data set by the search engines Mascot, OMSSA, X! Tandem, and their combination using OpenMS' ConsensusID and TPP's iProphet. In Figure S2 (Supporting Information), we show the number of identified spectra at different FDR levels (q-values[42]) for the tested methods. The q-values were computed with the FalseDiscoveryRate TOPP tool on the basis of a target-decoy approach. Generally, compared to the individual search engines, the number of identified spectra increased significantly using either of the combination approaches. For example, at the 1% FDR threshold (q-value 0.01) that we used in our analyses of label-free data, the combination approaches exceeded Mascot's identification rate by 54.4% (ConsensusID) and 57.4% (iProphet), respectively. OMSSA and X! Tandem were outperformed by ConsensusID (iProphet) by 23.2% (25.6%) and 37.0% (39.8%), respectively.

iProphet achieved higher identification rates than ConsensusID at q-values below 0.01, but the situation was reversed at q-values of 0.03 and above. Specifically, at a q-value of 0.01, both combination tools differed only marginally in the number of spectra they allowed to identify: Out of a total of 42 038 acquired MS2 spectra, ConsensusID facilitated peptide assignments for 12 785 spectra, while the iProphet approach assigned peptide sequences to 13 039 spectra. The difference of 254 spectra amounts to 2% of the identification totals.

On the peptide level, 2347 nonredundant peptides were jointly identified by both combination tools. Additionally, iProphet identified 349 unique peptides not found by ConsensusID, while ConsensusID assigned 174 peptide sequences not observed with iProphet.

Comparison of OpenMS Algorithms

We have introduced new algorithms for the TOPP tools PeakPicker, FeatureFinder, MapAligner, and FeatureLinker in this work. To compare the performance of the new algorithms to their established counterparts in OpenMS, we applied both old and new algorithms to the "Streptococcus" data set. We used the algorithms as implemented in OpenMS 1.8 in all cases; for the new algorithms, we applied the parameter settings as listed in Table S1 (Supporting Information), unless otherwise indicated. We did not test a second FeatureFinder algorithm, because OpenMS 1.8 does not provide an established alternative to the "centroided" algorithm: The FeatureFinder algorithms "simple" and "simplest" were removed as deprecated in OpenMS 1.7. The "isotope_wavelet" algorithm provides only experimental support for data from high-resolution instruments, and it would require special hardware (graphics processing units) to achieve acceptable runtimes on a large high-resolution data set.[43]

PeakPicker. In addition to the "high_res" algorithm, we applied the "wavelet" algorithm to all 58 raw data files from the "Streptococcus" data set. For the "wavelet" algorithm, we selected a signal-to-noise cutoff of zero (same as for the "high_res" algorithm), enabled the automatic estimation of the peak width, and kept all other parameters at their defaults. In our test, centroiding all the files with the "wavelet" algorithm took 4.3 times as long as with the "high_res" algorithm (6.6 vs 1.5 h), including reading and writing of the files. After centroiding, we applied the "centroided" FeatureFinder to the results of both algorithms. Since the goal in our label-free pipeline is to detect features in the data, we evaluate the quality of the centroiding algorithms based on the number of subsequently detectable features. Overall, the "high_res" results gave rise to 1.4 times as many features as the "wavelet" results (829 146 vs 587 166). This difference is mainly due to the "plasma proteins" experiment, where in "high_res"-centroided data, 2.6 times as many features as in "wavelet" results were detected. In the other two experiments, both centroiding algorithms enabled detection of comparable numbers of features, with a slight advantage for the "wavelet" algorithm (on average, 99/96% as many features found in "high_res" data

dx.doi.org/10.1021/pr300992u | J. Proteome Res. 2013, 12, 1628−1644

Table 1. Overview of the Tested Software Solutions for Label-Free Quantification: OpenMS, MaxQuant, and Progenesis LC–MS[a]

		OpenMS	MaxQuant	Progenesis
Characteristics	Availability	free and open-source	free	commercial
	Operating System	Windows, Mac OS, Linux	Windows	Windows
	Protein ID engines	various (external)	Andromeda (built-in)	various (external)
Coverage (peptides per sample)	"Dilution series"	1710 (FP: 166)	1517 (FP: 101)	2381 (FP: 1164)
	"Leptospira"	3111	3511	2914
	"Streptococcus"	3868	3798	3923
Accuracy (correlation)	"Dilution series"	0.84	0.90	0.80
	"Leptospira"	0.91 (FN: 13)	0.87 (FN: 11)	0.77 (FN: 6)

[a]Coverage (see also Figure 5) and accuracy (see Figures 6 and 7) are listed for the different data sets. Coverage was measured as the number of peptides that were quantified per sample by the full pipeline. FP: False positives, number of peptides that were quantified at 0% concentration (average of two samples). Accuracy in the "dilution series" data set was measured as the correlation between normalized peptide abundances and expected concentrations (average over all peptides that were quantified in at least four samples). Accuracy in the "Leptospira" data set was measured as the correlation of log-transformed abundances of all quantified proteins to the gold-standard values (average over three conditions). FN: False negatives, total number of gold-standard proteins that were not quantified in a sample (out of 117). There are no accuracy values for the "Streptococcus" data set because of the lack of a ground truth.

of the "dose response"/"fatty acids" experiment). However, for stricter parameter settings in FeatureFinder (higher score thresholds and lower m/z tolerances), we achieved better results with the "high_res" algorithm for all runs in this data set (data not shown). The increased reliability make the "high_res" algorithm our preferred choice for the preprocessing of high-resolution data.

MapAligner. In order to compare the new "identification" MapAligner algorithm to the previously described "pose_clustering" algorithm, we applied either algorithm to the three consensus maps from the subexperiments in the "Streptococcus" data set. This is a challenging test case because the LC–MS measurements for the three experiments were acquired several months apart, and because the "dose response" experiment used a longer chromatographic gradient than the other two experiments. For the latter reason, we chose the "dose response" consensus map as reference against which to align the other two maps. The "pose_clustering" algorithm, which we used with default parameters, required 15 min to compute the RT adjustment in our test (not including the necessary conversion of consensus maps to feature maps). In the parameters of the "identification" algorithm, we specified to only use peptides that were associated with features in all three consensus maps, and we disabled the "maximum RT shift" filter. The "identification" algorithm finished the calculation in only 75 s (no format conversion needed). The memory requirements of both algorithms were modest (below 2 GB).

We evaluate the RT adjustment results on the basis of all peptide identifications that the reference map shares with at least one of the other maps. While this type of information (for a subset of the peptides) was already utilized to compute the "identification" alignment, it provides the best standard for judging the alignment quality; for this reason, it was also used in the original validation of the "pose_clustering" algorithm.[36] The alignment results are depicted in Figure 3. They show that the "identification" algorithm was able to correct for most of the systematic differences in the peptide retention times, leaving just the intrinsic variation and some unresolvable remnants in the beginning and end of the RT scale. The latter are caused by peptides that were only separated by the longer gradient of the reference experiment; we cannot resolve such cases because our transformations map an original RT always to one value, not several. In contrast, results of the "pose_clustering" algorithm were impaired by its limitation to an affine transformation of the retention times. Fitting a B-spline to the feature landmarks identified by the "pose_clustering" algorithm would be possible, but not very useful, since an affine transformation (computed by pose clustering) is used to find those landmarks in the first place. So, although the "pose_clustering" alignment drastically reduced RT differences to the reference, it left significant deviations because nonlinear shifts could not be accounted for. Accordingly, the average absolute difference (between median RTs in the reference run and in the aligned runs) for the 6245 peptides in the test case is 94.1 s for the "identification" alignment, but 199.5 s for the "pose_clustering" alignment.

FeatureLinker. We compared the "unlabeled" and "unlabeled_qt" FeatureLinker algorithms by applying both to the consensus maps of the three experiments in the "Streptococcus" data set, after adjusting the maps to the same RT scale with MapAligner. The three maps contained a total of 171 535 consensus features. We applied the same settings for both algorithms; the additional grouping constraint provided by the "second_nearest_gap" parameter of the "unlabeled" algorithm (for which there is no equivalent in "unlabeled_qt") was disabled by setting the parameter to "1".

Given constraints for how similar features have to be to belong to the same group, a feature linking algorithm should produce feature groups that are as complete as possible (ideally containing one feature from each input map), or equivalently minimize the total number of groups needed to contain all features. The "unlabeled_qt" algorithm produced a solution with 129 261 feature groups, among them 11 305 complete groups of size three, in a runtime of about 2.7 h. In contrast, the result of the "unlabeled" algorithm comprised 133 790 groups overall (3.5% more), containing 6505 complete groups (42% less). Because of the very high number of (consensus) features per input map and the lack of optimizations in the "unlabeled" algorithm, that algorithm took a prohibitive 19.3 h to compute the feature grouping in our test. The memory requirements of both algorithms were similar, with around 2.5 GB each. On the basis of the different strategies for dealing with multiple inputs in the two algorithms, runtimes of the "unlabeled" algorithm will be more favorable compared to "unlabeled_qt" for higher numbers of input maps with fewer features in each map. However, in such cases, the qualitative difference of the

dx.doi.org/10.1021/pr300992u | J. Proteome Res. 2013, 12, 1628–1644

achieved grouping solutions will become even greater, favoring the more comprehensive clustering approach pursued in the "unlabeled_qt" algorithm. Generally, the "unlabeled_qt" algorithm will be able to produce a better overall solution than the "unlabeled" algorithm in all cases with more than two inputs.

Evaluation of the OpenMS Label-Free Pipeline

In this section, we discuss the overall performance of our pipeline and the interplay of its parts. We first look into the efficacy of the software with respect to detecting features that are relevant for the quantification of peptides and proteins. We then examine how the alignment of related runs helps to increase the fraction of peptides and proteins that can be quantified in each run. Finally, we compare the results produced by the full OpenMS pipeline to those of MaxQuant and Progenesis. We evaluate the performance of these methods based on two aspects: First, coverage: how many peptides/proteins could be quantified? And second, accuracy: how close are the quantification results to the expected values given by the ground truths in the "dilution series" and "Leptospira" data sets? An overview of the comparison is given in Table 1. In our assessments, we focus mostly on the peptide level, since we get a more complete picture before taking the many-to-many relationships between peptides and proteins into account. All numbers reported for identified peptides in the following are at 1% FDR and include decoys.

Feature Detection Performance. To assess the quality of the feature detection in our label-free data analyses, we examine the mapping of high-confidence peptide IDs to features. Since the reliable identification of a peptide sequence from an MS2 spectrum strongly suggests that the respective peptide is indeed present at that point in the LC−MS map, we expect to find a corresponding feature in the MS1 data. Peptide IDs without accompanying features thus indicate limitations in the feature detection process (false negatives). However, features that cannot be annotated with a peptide ID do not generally constitute false positives, because of the limited number of MS2 spectra that are typically acquired in an LC−MS run. It may even be possible to indirectly annotate some of these features after performing an alignment of related LC−MS runs (see next subsection). Finally, if a feature is annotated with peptides of different sequences (ambiguous annotation), this generally indicates an error, a misidentified peptide, a mistake in the feature detection, or an annotation mismatch. Features with ambiguous annotations do not contribute to the quantification. Table S2A (Supporting Information) lists summary statistics for different aspects of feature detection, peptide identification, and feature annotation in each of our data sets.

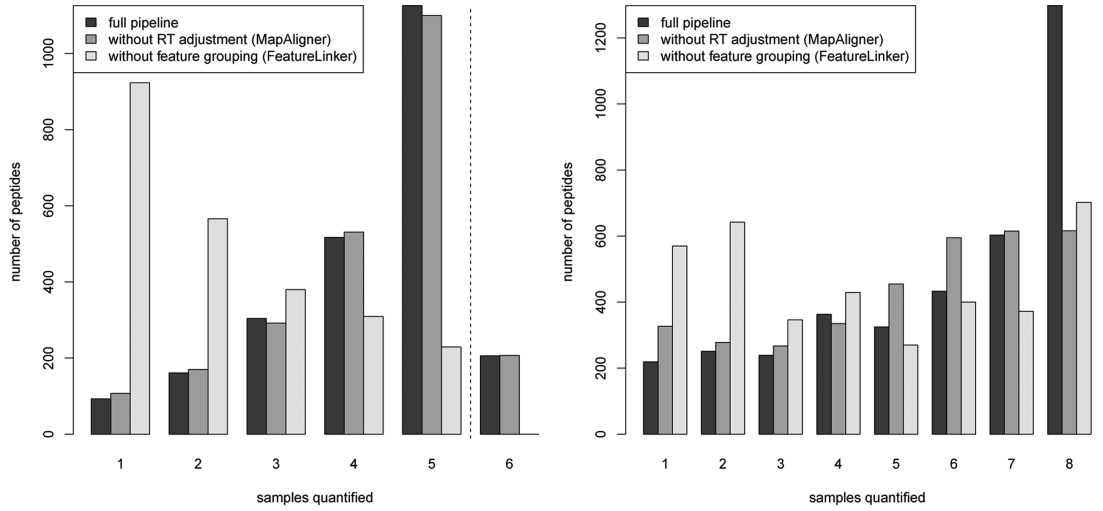
Processing the "dilution series" data set with the OpenMS pipeline, we detected on average 16 626 peptide features in each of the six runs. A higher number of features were found in runs with a higher fraction of human sample, in agreement with the fact that the human sample is of higher complexity than the *S. pyogenes* sample. With the TPP/multisearch engine approach, we obtained a relatively low number of peptide identifications per run in this data set, 2092 on average. Of these IDs, an average of 81% could be assigned to features by IDMapper. On average, only 6.7% of features detected in a run were annotated with one or more peptide IDs. About 0.6% of all annotated features exhibited an ambiguous annotation and could not be used for quantification. Considering only unique peptides with different sequences, 1057 were identified per individual run on average; over all runs, 2617 unique peptides were identified. Between 87 and 93% (average: 89%) of the unique peptides observed in an individual run could be quantified directly in that run. This translates to 33−39% (average: 36%) of the unique peptides identified in all runs together.

In the eight runs of the "Leptospira" data set, the TPP/multisearch engine approach identified on average 3761 MS2 spectra. Between 1686 and 1809 (average: 1749) unique peptides were identified per run, for a total of 4669 unique peptides in all runs together. These peptides map to 1431 different proteins or protein groups. Feature detection with OpenMS produced on average 16 210 features per run. We annotated each feature map with the combined peptide identifications from both related technical replicates. On average, 72% of the peptide IDs could be mapped to features. On the level of unique peptides, this corresponds to ratios between 63 and 67% (average: 66%) of the unique peptides in the relevant set of identifications, or on average 44% of all unique peptides identified. Again, only a small portion of the detected features (13% on average) could be associated with peptides at this stage. About 3% of all features annotated with peptides contained ambiguous annotations, precluding their use for quantification.

In the "Streptococcus" data set, the average number of features detected per run in the three subexperiments was 15 412 ("plasma proteins"), 16 924 ("dose response"), and 10 975 ("fatty acids"), respectively. In the "fatty acids" data, the number of detected features varied widely, but the variation was mostly consistent with the different growth conditions of the samples. The TPP searches yielded on average 6123 (1,379, "plasma proteins"), 13 459 (2942, "dose response"), and 4257 (2002, "fatty acids") peptide identifications (unique peptide identifications) per run. Here, variation was very high for the "plasma proteins" experiment because of the use of different data acquisition methods (DDA and two inclusion lists). By running IDMapper, an average of 79% ("plasma proteins"), 80% ("dose response"), and 87% ("fatty acids") of the peptide IDs could be annotated to features in each run of an experiment. On the level of unique peptides, on average 73% ("plasma proteins"), 89% ("dose response"), and 89% ("fatty acids") were directly quantifiable in their original run.

To test the effect of targeted feature detection, we applied the "centroided" FeatureFinder to the "Leptospira" data using seed lists derived from peptide IDs of the respective runs (via SeedListGenerator). We used tolerances of 120 s (RT) and 0.1 Da (m/z) for the filtering of internal seeds and kept the seed score threshold at 0.3. Compared to the standard usage of the algorithm, average runtimes were reduced by 95% (from 39 to 2 min) and average memory consumption was reduced by 80% (from 2.9 GB to 568 MB). At the same time, the number of detected features decreased by 80%, from an average of 16 210 to 3212 per run. However, the number of features in a run that could be associated with peptide IDs from the same run actually increased slightly, from an average of 1230 to 1252 (by about 2%). In this case, targeted feature detection thus provided superior performance, as long as the runs were considered individually. The downside of this approach is that potential gains from the alignment of multiple runs cannot be realized, since the additional features for which IDs might be inferred are largely missing.

Impact of the Alignment. In order to increase the utilization of features for the quantification of peptides, we aligned and combined the data from all runs in an experiment



(A) "Dilution series" dataset. (B) "Leptospira" dataset.

Figure 4. Impact of optional parts of the OpenMS label-free pipeline on the quantification coverage. Each bar in a plot represents the portion of peptides that could be quantified in the respective number of runs, using the indicated variant of the OpenMS pipeline. For the "full pipeline", we computed an alignment of the annotated feature maps by applying MapAligner followed by FeatureLinker and then performed quantification with ProteinQuantifier on the resulting consensus map. For the option "without RT adjustment", we used FeatureLinker directly and again quantified on the basis of the consensus map. "Without feature grouping" means we applied neither MapAligner (not useful in this case) nor FeatureLinker, performed quantification on the individual feature maps, and merged the results in postprocessing. (A) "Dilution series" data set. Bars at "6" indicate false positive matches in at least one of the runs (as neither human nor *S. pyogenes* peptides were present in all six samples, and almost no shared peptides were identified at 1% FDR). (B) "Leptospira" data set. Here, the variant "without RT adjustment" also implies that feature maps were only annotated with peptide identifications from the same run, not with combined IDs from both related technical replicates (as in the other two pipeline variants). The reason is that merging IDs from different runs without RT correction could lead to false positive assignments during feature annotation.

by applying the TOPP tools MapAligner and FeatureLinker. The resulting consensus maps contain groups of features from different runs. Features annotated with peptide IDs convey their annotations to the groups ("consensus features") that contain them; in turn, these annotations apply to all contained features. In this way, additional sequence annotations are inferred across runs for features not initially associated with a peptide ID. The impact of the alignment is thus determined by the increase in peptide quantifications that it provides. Summary statistics on the data set/experiment level, before and after the alignment, are listed in Table S2B (Supporting Information).

In the "dilution series" data set, the 99 759 features from all runs were grouped into 51 703 consensus features, each containing between one and six subfeatures. The 3223 consensus features (6.2%) that carried annotations contained 12 524 subfeatures. Comparing this to the total of 6528 features that could be directly annotated in the individual runs, the number of features contributing to the quantification could be almost doubled by performing the alignment. Accordingly, on the level of unique peptides, the fraction of the total 2617 peptides that could be quantified per run increased to between 46 and 77% (average: 65%). The beneficial effect of the alignment is also illustrated in Figure 4A. Ideally, all peptides would be quantified in all relevant runs. However, because of the setup of the dilution series, it is expected that some peptides may vanish beyond the detection limit at lower concentrations. Furthermore, almost no peptide should occur in all six runs, since one run contained purely *Streptococcus* sample and one purely human sample, and only three peptides common to both organisms were identified at the 1% FDR threshold. So, while the plot clearly shows that the feature grouping step significantly increases the quantification coverage, the bars for quantification in all six samples also indicate that some false positive matches were produced by FeatureLinker. On the "dilution series" data set, the adjustment of retention times with MapAligner prior to feature grouping provided only minimal benefit, because there were no big retention time shifts between the runs to begin with.

In the consensus map of the "Leptospira" data set, the 129 673 detected features were grouped into 45 445 consensus features, of which 4639 (10%) carried peptide annotations. The annotated consensus features conferred their identifications to 24 003 contained subfeatures. Counting only features with unambiguous annotation, performing the alignment increased the number of features available for quantification by 34%. At the same time, the number of unique peptides that could be quantified per run (on average) increased from 2080 to 2726, i.e., by 31%. On the protein level, this corresponds to a change from 851 to 993 unique proteins (or groups of proteins) per run that could be quantified on the basis of at least one peptide, an increase of 17%. The individual effect of the two alignment steps in the analysis of the "Leptospira" data set is illustrated in Figure 4B. The negative impact of forgoing RT adjustment with MapAligner was more severe here than in the "dilution series" case, because it also precluded us from merging peptide identifications from different inclusion list measurements of the same sample.

In the "Streptococcus" data set, a total of 8118 unique peptides were identified in all three experiments taken together. On the basis of individual runs alone, an average of 28% of those peptides could be quantified per DDA run (we exclude the inclusion list runs in the "plasma proteins" experiment from our consideration here, because the very low numbers of peptide identifications in them would skew the average). After performing the alignment of each individual experiment, the

dx.doi.org/10.1021/pr300992u | *J. Proteome Res.* 2013, 12, 1628−1644

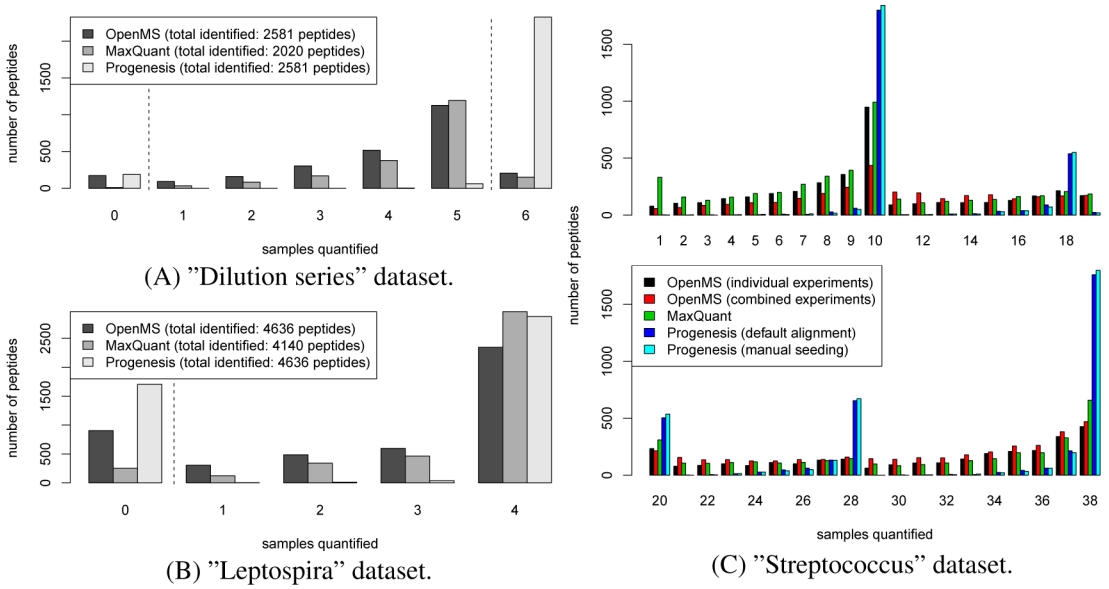


Figure 5. Quantification coverage on peptide level. The bars show how many peptides were quantified in how many runs. Bars at "0" represent identified peptides that were never quantified (A and B). Technical replicates were aggregated (B and C). (A) Results on the "dilution series" data set. Bars at "6" indicate false positive matches in at least one of the runs (as neither human nor *S. pyogenes* peptides were present in all six samples). (B) Results on the "Leptospira" data set. (C) Results on the "Streptococcus" data set. For MaxQuant (green bars) and Progenesis, results from the three subexperiments were merged in postprocessing. For OpenMS, both merging in postprocessing (black bars) and combination by alignment (red bars) are shown. For Progenesis, we distinguish between fully automatic alignments in each of the three experiments (blue bars) and alignments based on manual seeding (cyan bars).

number increased to 44%. By combining all experiments via a secondary alignment, we were able to achieve 49% average peptide quantification per run.

In summary, performing an alignment of multiple runs with the MapAligner and FeatureLinker tools in our tests increased the number of unique peptides that could be quantified per run considerably, by 30−80%. How much can be gained in a particular case is however difficult to predict, as this depends on a variety of factors, including the choice of alignment parameters, the size and complexity of the data set, the MS2 acquisition settings, and the prior processing of the data. In the case of large data sets composed of subgroups (individual experiments, patient/control groups, etc.), performing the alignment in a hierarchical fashion can help to mitigate differences in chromatography and to reduce the runtime/memory requirements of the computation.

Software Comparison: Coverage. When judging the overall performance of quantification methods that produce largely accurate results, how extensive those results are becomes an important factor. We call this aspect "coverage" and measure it in terms of peptide quantifications; i.e., we look at the numbers of distinct peptides quantified in each run/sample of a data set.

In Figure 5A, we compare the coverage that the three software solutions we tested on the "dilution series" data set. The plot clearly shows that Progenesis tends to aggressively match features across all runs, achieving high coverage at the cost of introducing false positives. We will also observe this effect below when we consider quantification accuracy. Altogether, Progenesis quantified 2328 peptides that map exclusively to *Streptococcus* or human proteins at 0% sample concentration of the respective proteome. In contrast, both OpenMS and MaxQuant produce relatively few such definitive mismatches (331 and 201 peptides quantified at 0%, respectively). Of all tools, MaxQuant provides the best coverage of its set of identified peptides. However, since this set is almost 22% smaller than that obtained by the TPP/multisearch engine approach, the total number of peptide quantifications is actually smaller for MaxQuant (9101, ∼1517 peptides per run) than for OpenMS (10 261, ∼1710 peptides per run).

The respective coverage achieved by OpenMS, MaxQuant, and Progenesis on the "Leptospira" data set is shown in Figure 5B. Since MaxQuant aggregated results from the technical replicates in its output, we have done the same for OpenMS and Progenesis in postprocessing. A peptide counts as "quantified" in a biological sample if it was quantified in at least one of the two corresponding replicate runs. Overall, MaxQuant achieved the best coverage, producing altogether 14 045 quantitative values for peptides in all four biological samples (∼3511 peptides quantified per sample). OpenMS comes second with 12 444 peptide quantifications (∼3111 per sample). Progenesis follows (11 654 abundance values, ∼2914 per sample), again achieving complete coverage for almost all peptides it quantifies. We believe that the main cause for the diminished coverage of OpenMS on this data set is a property of the feature detection: The stringent feature model used may too easily reject features with unusual characteristics, for example, features that exhibit very jagged elution profiles, which is a common occurrence in the "Leptospira" data.

In Figure 5C, we illustrate the situation for the "Streptococcus" data set. Again, technical replicates in the "plasma proteins" experiment were aggregated during processing (MaxQuant) or postprocessing (OpenMSProgenesis), leaving us with 38 samples. With MaxQuant and Progenesis, it was only possible to quantify data in each of the three experiments individually, not to combine all three. Nevertheless, both methods achieved high coverage, even when considering the whole data set. MaxQuant produced a total of 144 316 quantitative values for peptides, i.e., quantified on

dx.doi.org/10.1021/pr300992u | J. Proteome Res. 2013, 12, 1628−1644

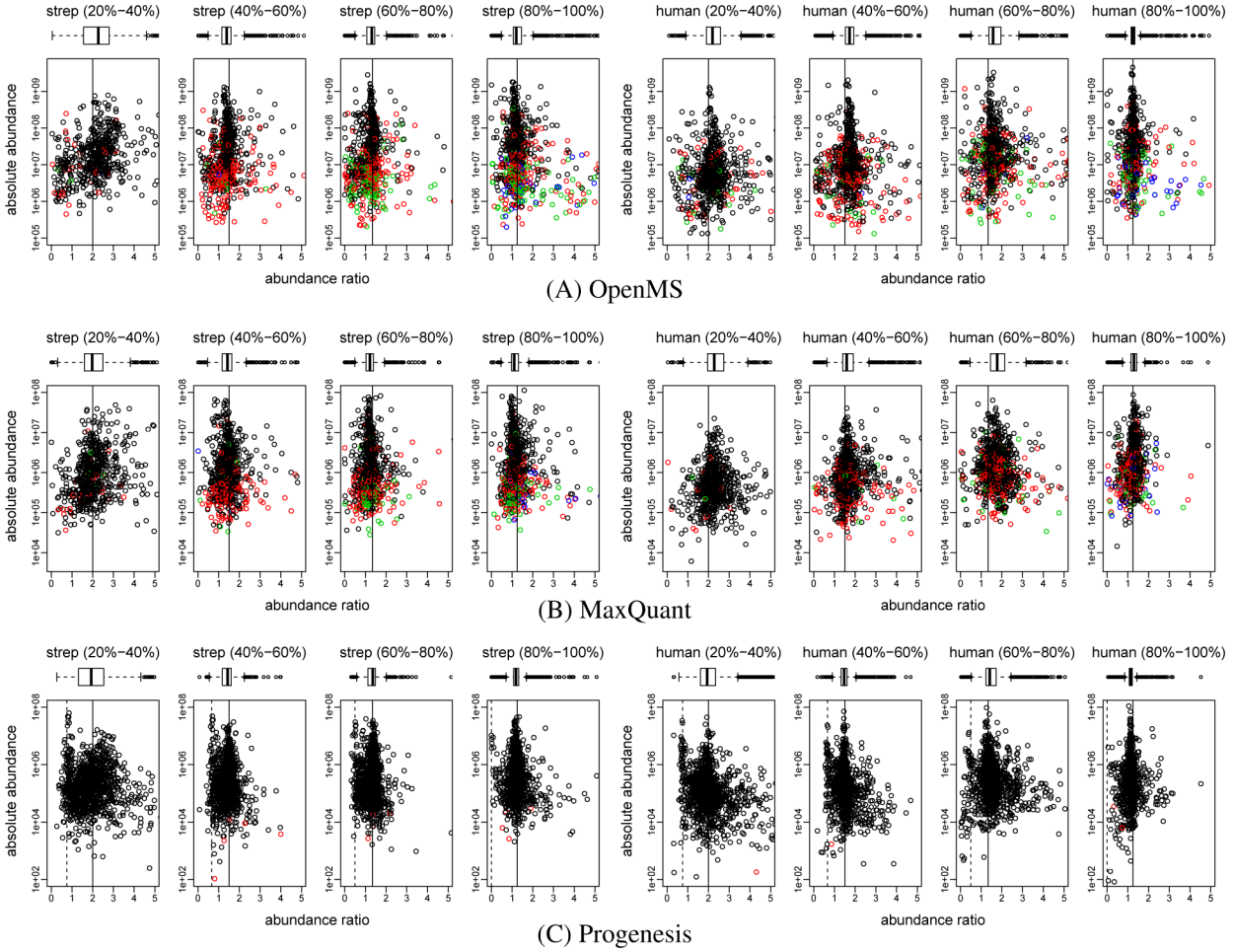


Figure 6. Quantification accuracy on peptide level for different software tools on the "dilution series" data set. Each individual plot shows normalized peptide abundances from two adjacent concentrations (top row: *S. pyogenes* peptides, bottom row: human peptides). Summed abundances from both concentrations are plotted against fold changes in abundance when going from the lower to the higher concentration. A box plot at the top shows the distribution of observed fold changes; a continuous vertical line indicates the theoretical value for the fold change (e.g., going from 20 to 40%, a 2-fold increase in abundance is expected). Points are colored according to the number of relevant samples in which a peptide was quantified (black: 5, red: 4, green: 3, blue: 2). The additional dashed vertical lines in the Progenesis plots indicate the fold changes that would be expected for mismatched peptides from the other organism in the same samples.

average 3798 peptides per sample. For Progenesis, we tested two variants of the alignment in each experiment: With "all automatic" alignments, 148 153 quantitative values (~3899 peptides per sample) were obtained. With manual seeding of the alignments (as in the "dilution series" and "Leptospira" data sets), the number could be increased to 149 061 (~3923 peptides per sample). Notably, Progenesis with manually seeded alignments quantified 1799 peptides in all 38 samples. For OpenMS, the peptide coverage was comparatively low when results from the three experiments were considered individually: 130 419 peptide quantifications were obtained, corresponding to 3432 peptides per sample on average. However, as the only software in our test, OpenMS allows to combine different experiments by performing a secondary alignment of the consensus maps. This enables the inference of additional peptide annotations for features across experiments, increasing the coverage. As a consequence, the final result of the OpenMS pipeline contained 147 001 quantitative values for peptides, or on average 3868 peptides per sample (2% more than MaxQuant, 1% less than Progenesis).

Software Comparison: Accuracy. Finally, we evaluate the accuracy of the quantification results generated by the OpenMS pipeline, MaxQuant, and Progenesis. We do this on the basis of the ground truths available for the "dilution series" and "Leptospira" data sets. Since there is no ground truth for the "Streptococcus" data set, we do not consider it here.

To assess quantification accuracy on the "dilution series" data set, we examined the ratios of normalized peptide abundances obtained for different sample concentrations. Knowing the concentrations of *Streptococcus* and human sample in each run, we also know what fold changes to expect when peptide abundances from two runs are compared. In Figure 6, we plot absolute abundances against fold changes for *Streptococcus* and human peptides in pairs of adjacent runs. In all cases, quantification accuracy was generally better for high-abundance peptides than for low-abundance ones. Accordingly, quantification results tended to become more accurate at higher concentrations. The quantification results of OpenMS and MaxQuant were overall similar, and in both cases, some peptides could only be quantified at higher concentrations. With Progenesis, however, almost all peptides were quantified in all runs, but not always correctly. Especially at lower sample concentrations, the Progenesis results exhibited a distinctive subpopulation in which peptides were apparently mis-

dx.doi.org/10.1021/pr300992u | *J. Proteome Res.* 2013, 12, 1628–1644

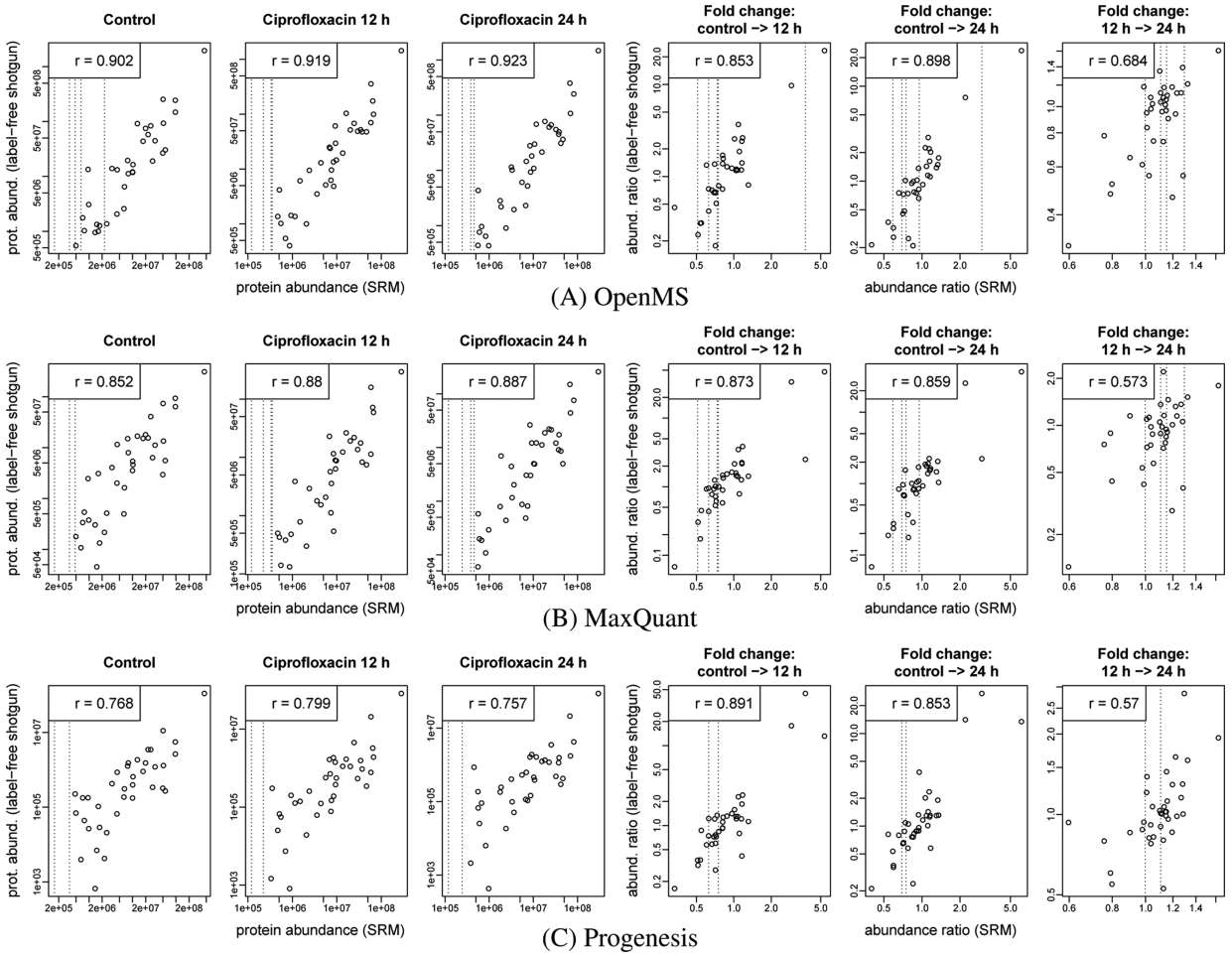


Figure 7. Quantification accuracy on protein level, comparing SRM and shotgun results of different software tools on the "Leptospira" data set. Shotgun results for the "control" condition are averages of the two biological replicates. Dotted vertical lines indicate missing values due to proteins that could not be quantified in the shotgun approach. Pearson correlation coefficients (r) of log-transformed abundance values/ratios are given in each plot.

quantified on the basis of data from the wrong organism. The fold changes for this subpopulation coincided with the values that we would expect for peptides from the other organism present in the same two samples (e.g., for human peptides at 20 and 40%, the expected fold change is 2; the same runs contain *Streptococcus* peptides concentrated at 80 and 60%, for an expected fold change of 0.75). As a measure of overall quality, we considered the average correlation between normalized peptide abundances and the scale of expected concentrations. For peptides that were quantified in at least four runs, we obtained the best result for MaxQuant (average correlation of 0.90 based on 1724 peptides), followed by OpenMS (0.84 based on 1847 peptides) and Progenesis (0.80 based on 2392 peptides). These numbers indicate a trade-off between coverage and overall quantification accuracy of the three tested methods.

To judge the accuracy of label-free quantification on the "Leptospira" data set, we compared our results to protein abundances computed from SRM measurements of the same samples (Figure 7). SRM data was available for 39 proteins, which spanned an abundance range of over 3 orders of magnitude. None of the label-free quantification approaches tested were able to quantify all 39 proteins; in fact, the two least abundant proteins were not even identified by any matching MS2 spectra in our searches of the shotgun data. Apart from these two, Progenesis quantified all proteins in the set. MaxQuant missed three to four low-abundance proteins of the 39 (depending on the sample), and OpenMS missed four to five. The proteins that were quantified with OpenMS cover a dynamic range of 2.7−3 orders of magnitude in the gold-standard SRM abundances, depending on the condition. Looking at the accuracy of the estimated protein abundances, measured in terms of correlation to the SRM-based values (Pearson correlation coefficient of logarithms of abundances), the ranking of the software tools is reversed: OpenMS performed best, with correlations over 0.9, followed by MaxQuant (correlations around 0.87) and Progenesis (around 0.78). The relations still apply if only proteins quantified by all software tools are considered; i.e., a higher correlation is not merely an artifact of failing to quantify "difficult" proteins. In the case of Progenesis, the agreement with SRM results increases if fold changes in protein abundance from the control condition to a stimulation are considered, instead of absolute protein abundances. However, since this improvement could not be observed when the two stimulations are compared, it may simply be due to a random effect. Taken together, our results on this data set again point to a trade-off between accuracy and coverage achieved by different methods.

dx.doi.org/10.1021/pr300992u | *J. Proteome Res.* 2013, 12, 1628−1644

DISCUSSION

In this study, we have presented an integrated processing pipeline for the label-free quantification of peptides and proteins, composed of tools and algorithms from the OpenMS software framework that were newly developed or systematically improved for this purpose. We have benchmarked our pipeline on the basis of three sets of LC−MS/MS experiments and have shown how the new algorithms improve upon their established counterparts in OpenMS, while the other enhancements increase the usability of the pipeline. We have further compared the performance of our software to that of the alternatives MaxQuant and Progenesis LC−MS.

We have found that all three software solutions produce adequate and largely comparable quantification results; all have some weaknesses, and none can outperform the other two in every aspect that we examined. However, the performance of OpenMS is on par with that of its two tested competitors while being open source, operating system independent, and highly flexible. Through the ability of the OpenMS label-free pipeline to interface with multisearch engine approaches for peptide and protein identification, we could make use of a wealth of high-confidence identifications that would otherwise not have been accessible for quantification. In each run of the "dilution series" data set ("Leptospira" data set, "Streptococcus" data set), we were able to quantify on average 65% (58%, 49%) of all unique peptides identified in the whole data set. This corresponds to 161% (155%, 216%) of the average number of unique peptides identified per individual run in the data set. Here, quantification coverage was aided by the new algorithms used to align and combine different runs. Even for low-abundance peptides and proteins, we achieved exemplary quantification accuracy with the OpenMS pipeline, as evidenced by the good agreement with expected fold changes in the "dilution series" data set, as well as by the unequaled correlations of 0.9 and above with the gold-standard SRM data in the "Leptospira" data set. In the case of the heterogeneous "Streptococcus" data set, OpenMS was the only software that could produce a joint result for all three subexperiments.

MaxQuant consistently performed well in our tests, but it would benefit from the ability to import peptide identifications from multi-search-engine approaches, in order to increase the number of peptides and proteins that could potentially be quantified. In accordance with previous results,[24,26,28] we have found that combining results from multiple search engines can significantly improve the identification rate at a given FDR level; at low FDRs (between 1 and 5%), we have observed equal gains from the two combination tools iProphet and ConsensusID. Compared to MaxQuant's Andromeda search engine, at 1% FDR the iProphet approach identified 27.8 and 13.4% more unique peptides on our two test data sets. As another downside, MaxQuant is only applicable to data from Thermo instruments.

Unlike MaxQuant, Progenesis has the ability to use results from iProphet searches (via import of a pepXML file). In our tests, it has achieved exceptional quantification coverage by aggressively matching features across different runs, but at the risk of incorporating a significant number of false positive matches in the process. This may negatively affect quantification accuracy in the end; however, the extent to which this happens depends on the data set. Progenesis users should thus take this effect into account and avoid processing dissimilar runs together in the same session.

Although we have focused on quantification performance in this work, there are other aspects to consider when evaluating software tools for label-free data analysis: In contrast to many other programs (like SuperHirn, MaxQuant, or Progenesis), OpenMS does not offer an integrated solution specifically for label-free quantification. Instead, with the TOPP tools OpenMS provides a wide selection of applications for LC−MS data processing, a subset of which can be chained into a label-free data analysis pipeline. This concept allows for great flexibility in tailoring a pipeline to different problem configurations. It also makes it easy to automate the data analysis to a large extent. On the downside, not offering a full processing pipeline "out of the box" means that users need to become familiar with the individual tools in order to be able to construct their own pipelines. We hope to reduce this obstacle with the present publication; also, TOPPAS workflows for ubiquitous processing tasks are being made available for download on the OpenMS Web site (http://openms.de/downloads/toppasworkflows). As another result of OpenMS' focus on openness and flexibility, some TOPP tools expose a large number of parameters that allow control over various facets of their behavior. Parameter optimization, especially on the level of whole processing pipelines, can thus become a daunting task. Extensive documentation, useful default values, and the designation of advanced parameters (that usually should not be changed) as such counteract this problem. In addition, there are attempts to intelligently estimate parameters for some tools from the data (e.g., in the "wavelet" PeakPicker algorithm).

As a further difference from MaxQuant or Progenesis, OpenMS does not provide a graphical interface to guide the user through the data analysis. With TOPPView[44] it however contains a powerful viewer for raw data in various file formats as well as derived data in OpenMS XML formats (featureXML, consensusXML, idXML). This allows users to visually evaluate processing results at all stages of the pipeline.

Finally, OpenMS does not include statistical tools for the evaluation of the quantification results. However, open file formats and the ability to export data to TSV/CSV files (TextExporter) would make it relatively easy for tools that provide such functionality, like Corra[45] or MSstats,[46] to make use of data produced by OpenMS.

Although OpenMS as a framework has a somewhat different orientation than the MaxQuant or Progenesis software suites, the algorithms and tools that it provides make it an effective solution for label-free quantification of LC−MS/MS data, one that achieves state-of-the-art quantification accuracy and coverage. Smart algorithms, the use of the TOPP concept of independent applications, and support for different operating systems, allowing OpenMS to run on the most powerful hardware available, enable us to process even large data sets (upward of 50 LC−MS/MS runs) with the OpenMS label-free pipeline. The open-source development model along with the dedication of the development team ensure that OpenMS will continue to improve and to provide innovative software in the future.

ASSOCIATED CONTENT

Supporting Information

Supplemental figures and tables. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: lars@imsb.biol.ethz.ch.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

For providing us with the data used in this work, the authors thank Johan Malmström ("dilution series" and "Streptococcus" data sets), Alexander Schmidt ("Leptospira" shotgun data), and Christina Ludwig ("Leptospira" SRM results). We are grateful to Bastian Blank for his contribution to the MaxQuant analyses. We also want to acknowledge all of the developers who contributed to OpenMS, in particular Steffen Sass for work on the "unlabeled_qt" FeatureLinker algorithm. Finally, we thank the reviewers for helpful comments.

REFERENCES

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422*, 198−207.
(2) Eng, J. K.; McCormack, A. L.; Yates, J. R. 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.
(3) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551−3567.
(4) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3*, 958−964.
(5) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466−1467.
(6) Tanner, S.; Shu, H.; Frank, A.; Wang, L.-C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **2005**, *77*, 4626−4639.
(7) Tabb, D. L.; Fernando, C. G.; Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **2007**, *6*, 654−661.
(8) 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.* **2002**, *74*, 5383−5392.
(9) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923−925.
(10) Domon, B.; Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.* **2010**, *28*, 710−721.
(11) Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **2007**, *389*, 1017−1031.
(12) Liu, H.; Sadygov, R. G.; Yates, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **2004**, *76*, 4193−4201.
(13) Lange, V.; Picotti, P.; Domon, B.; Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **2008**, *4*, 222.
(14) Ludwig, C.; Claassen, M.; Schmidt, A.; Aebersold, R. Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry. *Mol. Cell. Proteomics* **2012**, *11*, M111.013987.
(15) Schiess, R.; Wollscheid, B.; Aebersold, R. Targeted proteomic strategy for clinical biomarker discovery. *Mol. Oncol.* **2009**, *3*, 33−44.
(16) Schmidt, A.; Beck, M.; Malmström, J.; Lam, H.; Claassen, M.; Campbell, D.; Aebersold, R. Absolute quantification of microbial proteomes at different states by directed mass spectrometry. *Mol. Syst. Biol.* **2011**, *7*, 510.
(17) Mueller, L. N.; Rinner, O.; Schmidt, A.; Letarte, S.; Bodenmiller, B.; Brusniak, M.-Y.; Vitek, O.; Aebersold, R.; Müller, M. SuperHirn—a novel tool for high resolution LC−MSbased peptide/protein profiling. *Proteomics* **2007**, *7*, 3470−3480.
(18) 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.
(19) Sturm, M.; Bertsch, A.; Gröpl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O. OpenMS—an open-source software framework for mass spectrometry. *BMC Bioinf.* **2008**, *9*, 163.
(20) Kohlbacher, O.; Reinert, K.; Gröpl, C.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Sturm, M. TOPP—the OpenMS proteomics pipeline. *Bioinformatics* **2007**, *23*, e191−e197.
(21) Malmström, J.; Karlsson, C.; Nordenfelt, P.; Ossola, R.; Weisser, H.; Quandt, A.; Hansson, K.; Aebersold, R.; Malmström, L.; Björck, L. *Streptococcus pyogenes* in human plasma: adaptive mechanisms analyzed by mass spectrometry-based proteomics. *J. Biol. Chem.* **2012**, *287*, 1415−1425.
(22) Gillespie, J. J.; et al. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect. Immun.* **2011**, *79*, 4286−4298.
(23) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007**, *7*, 655−667.
(24) Quandt, A.; Espona, L.; Balasko, A.;Weisser, H.; Brusniak, M.-Y.; Kunszt, P.; Aebersold, R.; Malmström, L. *J. Proteome Res.*, submitted for publication.
(25) Deutsche, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; Eng, J. K.; Martin, D. B.; Nesvizhskii, A. I.; Aebersold, R. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10*, 1150−1159.
(26) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **2011**, *10*, M111.007690.
(27) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, 4646−4658.
(28) Nahnsen, S.; Bertsch, A.; Rahnenführer, J.; Nordheim, A.; Kohlbacher, O. Probabilistic consensus scoring improves tandem mass spectrometry peptide identification. *J. Proteome Res.* **2011**, *10*, 3332−3343.
(29) Martens, L.; et al. mzML—a community standard for mass spectrometry data. *Mol. Cell. Proteomics* **2011**, *10*, R110.000133.
(30) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011**, *10*, 1794−1805.
(31) Junker, J.; Bielow, C.; Bertsch, A.; Sturm, M.; Reinert, K.; Kohlbacher, O. TOPPAS: a graphical workflow editor for the analysis of high-throughput proteomics data. *J. Proteome Res.* **2012**, *11*, 3914−3920.
(32) Lange, E.; Gröpl, C.; Reinert, K.; Kohlbacher, O.; Hildebrandt, A. High-accuracy peak picking of proteomics data using wavelet techniques. *Pac. Symp. Biocomput.* **2006**, 243−254.
(33) Sturm, M. Ph.D. thesis, Universität Tübingen: Tübingen, Germany, 2010.
(34) Senko, M. W.; Beu, S. C.; McLafferty, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 229−233.

(35) Lan, K.; Jorgenson, J. W. A hybrid of exponential and gaussian functions as a simple model of asymmetric chromatographic peaks. *J. Chromatogr., A* **2001**, *915*, 1−13.
(36) Lange, E.; Gröpl, C.; Schulz-Trieglaff, O.; Leinenbach, A.; Huber, C.; Reinert, K. A geometric approach for the alignment of liquid chromatography-mass spectrometry data. *Bioinformatics* **2007**, *23*, i273−i281.
(37) Fischer, B.; Roth, V.; Buhmann, J. M. Time-series alignment by non-negative multiple generalized canonical correlation analysis. *BMC Bioinf.* **2007**, *8* (Suppl 10), S4.
(38) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **2006**, *78*, 779−787.
(39) Heyer, L. J.; Kruglyak, S.; Yooseph, S. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.* **1999**, *9*, 1106−1115.
(40) Silva, J. C.; Gorenstein, M. V.; Li, G.-Z.; Vissers, J. P. C.; Geromanos, S. J. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **2006**, *5*, 144−156.
(41) Malmström, J.; Beck, M.; Schmidt, A.; Lange, V.; Deutsch, E. W.; Aebersold, R. Proteomewide cellular protein concentrations of the human pathogen *Leptospira interrogans*. *Nature* **2009**, *460*, 762−765.
(42) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Posterior error probabilities and false discovery rates: two sides of the same coin. *J. Proteome Res.* **2008**, *7*, 40−44.
(43) Hussong, R.; Gregorius, B.; Tholey, A.; Hildebrandt, A. Highly accelerated feature detection in proteomics data sets using modern graphics processing units. *Bioinformatics* **2009**, *25*, 1937−1943.
(44) Sturm, M.; Kohlbacher, O. TOPPView: an open-source viewer for mass spectrometry data. *J. Proteome Res.* **2009**, *8*, 3760−3763.
(45) 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.; Watts, J. D. Corra: Computational framework and tools for LC−MS discovery and targeted mass spectrometry-based proteomics. *BMC Bioinf.* **2008**, *9*, 542.
(46) Clough, T.; Key, M.; Ott, I.; Ragg, S.; Schadow, G.; Vitek, O. Protein quantification in labelfree LC−MS experiments. *J. Proteome Res.* **2009**, *8*, 5275−5284.