Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses

George Rosenberger1,2,12, Isabell Bludau1,2,12, Uwe Schmitt3, Moritz Heusel1,4, Christie L Hunter5, Yansheng Liu1, Michael J MacCoss6, Brendan X MacLean6, Alexey I Nesvizhskii7,8, Patrick G A Pedrioli1, Lukas Reiter9, Hannes L Röst1, Stephen Tate10, Ying S Ting6, Ben C Collins1,6 & Ruedi Aebersold1,11,12

Liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) is the main method for high-throughput identification and quantification of peptides and inferred proteins. Within this field, data-independent acquisition (DIA) combined with peptide-centric scoring, as exemplified by the technique SWATH-MS, has emerged as a scalable method to achieve deep and consistent proteome coverage across large-scale data sets. We demonstrate that statistical concepts developed for discovery proteomics based on spectrum-centric scoring can be adapted to large-scale DIA experiments that have been analyzed with peptide-centric scoring strategies, and we provide guidance on their application. We show that optimal tradeoffs between sensitivity and specificity require careful considerations of the relationship between proteins in the samples and proteins represented in the spectral library.

We propose the application of a global analytic constraint to prevent the accumulation of false positives across large-scale data sets. Furthermore, to increase the quality and reproducibility of published proteomic results, well-established confidence criteria should be reported for the detected peptide queries, peptides and inferred proteins.

Technological advances in LC-MS/MS have greatly advanced our capabilities to explore proteomes. In bottom-up proteomics, the most widely used approach, proteins are proteolytically digested into peptides to increase their accessibility by LC-MS/MS. These peptides are then ionized and processed to generate fragment ion spectra (i.e., MS/MS spectra) that can be used to derive the amino acid sequences.

Several classes of bottom-up proteomic methods have been developed that differ in the way the peptide ions are selected for fragmentation and for how the resulting spectra are processed computationally. Currently, three main data acquisition strategies are applied: data-dependent acquisition (DDA), targeted acquisition by selected or parallel reaction monitoring (SRM or PRM, respectively) and DIA. Each class of methods has specific strengths and weaknesses, which have been extensively discussed1–3. The acquired data can be analyzed by different strategies, and the two main approaches differ in their query unit and are referred to as spectrum-centric and peptide-centric scoring methods4. In spectrum-centric scoring approaches, which have been implemented for DDA and DIA5–8 data analysis, a spectrum or pseudospectrum (when generated from DIA data) is queried against a peptide-sequence database to determine the peptide sequence that best matches the spectrum. In peptide-centric scoring methods, which have mainly been applied to SRM, PRM or DIA9–13 data, a peptide of interest is queried with specific peptide-query parameters against the data to find the best candidate peptide signal(s)4.

Peptide-query parameters are also referred to as ‘transition lists’ or ‘Tier 3’ assays14, which include sets of precursor and product ion m/z pairs that, in combination, enable selective and sensitive detection of a peptide by a ‘peak group’ of co-eluting fragment ion chromatograms (Supplementary Table 1).

Although these signal processing and scoring systems are applicable to data sets of varying size and complexity, special attention needs to be paid to appropriate methods of error-rate control to prevent the accumulation of false-positive identifications, particularly in cases in which large sample cohorts are analyzed. The false discovery rate (FDR)15 is a metric used for controlling the error rate of identified or detected analytes in experiments that are affected by the multiple-testing problem. It is currently the most commonly used metric within the field of mass-spectrometry-based proteomics. For proteomic analyses, thresholding is often applied to the raw confidence values to identify peptides that are confidently identified. Ideally, these confidence levels should be reported for the detected peptide queries, peptides and inferred proteins.

Two main approaches differ in their query unit and are referred to as spectrum-centric and peptide-centric scoring methods. In spectrum-centric scoring approaches, a spectrum or pseudospectrum is queried against a peptide-sequence database to determine the peptide sequence that best matches the spectrum. In peptide-centric scoring methods, a peptide of interest is queried with specific peptide-query parameters against the data to find the best candidate peptide signal(s). Peptide-query parameters are also referred to as ‘transition lists’ or ‘Tier 3’ assays, which include sets of precursor and product ion m/z pairs that, in combination, enable selective and sensitive detection of a peptide by a ‘peak group’ of co-eluting fragment ion chromatograms.

1Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland. 2PhD Program in Systems Biology, University of Zurich and ETH Zurich, Zurich, Switzerland. 3ID Scientific IT Services, ETH Zurich, Zurich, Switzerland. 4PhD program in Molecular and Translational Biomedicine, Competence Center Personalized Medicine (CC-PM), ETH Zurich and University of Zurich, Zurich, Switzerland. 5SIEX, Redwood City, California, USA. 6Department of Genome Sciences, University of Washington, Seattle, Washington, USA. 7Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA. 8Department of Pathology, University of Michigan, Ann Arbor, Michigan, USA. 9Biognosys, Schlieren, Switzerland. 10SIEX, Concord, Ontario, Canada. 11Faculty of Science, University of Zurich, Zurich, Switzerland. 12These authors contributed equally to this work. Correspondence should be addressed to R.A. (aebersold@imsb.biol.ethz.ch) or B.C.C. (collins@imsb.biol.ethz.ch).

RECEIVED 12 SEPTEMBER 2016; ACCEPTED 7 JULY 2017; PUBLISHED ONLINE 21 AUGUST 2017; DOI:10.1038/NMETH.4398
proteomics and can be estimated by different methods, including by derivation from posterior error probabilities that are estimated without or with the help of decoys or by using nonparametric \( q \)-value estimation by the target-decoy approach. Conversely, the false nondiscovery rate (FNR)\(^{19-21} \), which represents the rate of ‘missed’ discoveries at selected thresholds, provides a controllable metric for sensitivity under the same assumptions as those for the FDR (Supplementary Note 1).

Error-rate control originally emerged as a critical issue in DDA-based discovery proteomics as a result of advances in experimental design and instrumentation that generated data sets of increasing size\(^{22} \). Peptide identification is generally established by spectrum-centric database searches, and statistical modeling provides error-rate control at different levels, including peptide–spectrum matches (PSMs)\(^{16,23} \) and inferred proteins\(^{24,25} \), and for different experimental contexts\(^{26-29} \). Although the underlying algorithms for error-rate control are different, their results have been shown to converge within the boundaries of their assumptions\(^{18,30} \).

In contrast, targeted proteomic methods are commonly used in cases for which specific peptides need to be quantified across large sample cohorts with a high degree of reproducibility and quantitative accuracy\(^{31} \). In this type of measurement, it is expected that the majority of targeted peptides will be detected in most of the samples, thus reducing the detection challenge mainly to that of selecting and quantifying the correct peptide fragment signals, which are also referred to as ‘peak groups’\(^{32} \). Data generated by SRM or PRM targeted proteomic measurements are therefore not affected by the same statistical challenges as typical spectrum-centric discovery proteomics experiments.

Recent developments in MS technology, specifically the development of DIA methods\(^{2,3} \) and their application to cohorts consisting of hundreds of samples, have led to the generation of complex data sets, which consist of large numbers of measured peptides (typically thousands to tens of thousands per sample), the presence and quantity of which need to be established and compared over many samples. This presents challenges for peptide-level and protein-level error-rate control in peptide-centric analysis of DIA data, particularly in cases in which comprehensive spectral libraries (i.e., those that cover a substantial fraction of the proteome)\(^{33-36} \), are being used\(^{36} \). Such analyses involve tens of thousands of peptide queries per sample across tens to hundreds of runs, which can lead to substantial error accumulation when the resulting multiple-testing challenges are not addressed appropriately.

Here we propose that the criteria established for confidence assessment (i.e., based on multiple-testing-corrected error-rate estimation) of identified peptides and inferred proteins in spectrum-centric analysis should also be applied to peptide-centric scoring methods at both the peptide and protein levels for such studies. We show that data interpretation is dependent on the experimental context and offer considerations for designing an optimal analysis strategy. We demonstrate the applicability of the described concepts on a SWATH-MS interlaboratory reproducibility study\(^{37} \) and a human blood plasma data set comprising hundreds of samples\(^{38} \). We discuss the tradeoffs between spectral library comprehensiveness and sample specificity in light of their respective requirements for appropriate error-rate control.

**RESULTS**

**Peptide queries based on sample-specific versus combined spectral libraries**

Most of the published studies using DIA with targeted data extraction have used sample-specific spectral libraries that were generated either from corresponding DDA runs\(^{9,10,13,38-41} \) or from the DIA data itself. When sample-specific spectral libraries are used, it is usually sufficient to perform error-rate control on the peptide-query level only. Because the content of spectral libraries is restrictively filtered during the process of generating the library\(^{42} \), putative false-positive proteins are unlikely to be included in the targeted data extraction step (Supplementary Note 2). This is not the case for spectral libraries that are generated from multiple heterogeneous samples, for example, different cell or tissue types. In such cases, the spectral library contains a large fraction of ‘false targets’ that are not detectable in a specific sample. This value is also referred to as \( \pi_0 \) (ref. 43). The \( \pi_0 \) value is directly coupled to the error estimation within a data set, where larger \( \pi_0 \) values require stricter multiple-hypothesis testing, as well as strategies to control for error accumulation from the PSM- or peptide-query level to the protein level, as has been demonstrated for discovery proteomics\(^{37,44} \). This effect is further accentuated when repository-scale spectral libraries, such as our combined human assay library (CAL)\(^{36} \), are used to analyze large sample cohorts.

In light of these considerations, the ideal case for most scenarios would be peptide queries that exactly match the set of detectable targets in the DIA data set. However, comprehensive libraries can substantially increase the sensitivity of peptide-centric scoring approaches\(^{36,37} \) and are required to quantitatively compare heterogeneous samples in larger-sized cohorts, such as those in clinical studies (Supplementary Note 3). Thus, it is crucial to apply robust error-rate control methods in workflow schemes for peptide-centric scoring, particularly in cases of high \( \pi_0 \) values, similar to that for the situation in discovery proteomics.

**Protein FDR assessment**

As stated above, error-rate control on only the peptide-query level is insufficient to infer sets of proteins in workflows that use comprehensive spectral libraries that result in high \( \pi_0 \) values. For these cases, we previously suggested that the error rate should be controlled not only on the peptide-query level but also on the peptide and protein levels\(^{36} \). This can be achieved, for example, by adapting
a target-decoy approach as that initially implemented for protein-level spectrum-centric analyses in MAYU\textsuperscript{27} or, more recently, SWATH2stats\textsuperscript{45}. Another option is the application of nonparametric modeling strategies to compute posterior probabilities at the peptide and protein levels\textsuperscript{17}, as those that have been adapted for DIA analyses in DIA-Umpire\textsuperscript{6,66} and SWATHProphet\textsuperscript{47}.

Alternatively, the \(q\)-value\textsuperscript{33} has been proposed for error estimation at the PSM level, as well as at the protein level\textsuperscript{18}. The \(q\)-value is a measure of significance for analyte detection that is comparable to the \(p\)-value, except that it accounts for multiple-testing in a manner analogous to the FDR. Here we have investigated whether peptide-level and protein-level \(q\)-values could be estimated similarly to the peptide-query level \(q\)-values in our workflow consisting of OpenSWATH\textsuperscript{10} and PyProphet\textsuperscript{11}, a re-implementation of the mProphet\textsuperscript{32} algorithm for DIA data. Whereas OpenSWATH and related tools compute a set of scores for each peptide query, PyProphet combines these scores to a single discriminant score by applying semi-supervised learning to best separate decoys from authentic peptides.

Figure 2 | Schematic illustration of the different context-dependent error-rate estimation strategies. Top, the ‘run-specific context’ conducts separate \(q\)-value estimation for each sample. This method results in run-specific \(q\)-values that can represent different peak group qualities between runs with varying \(\pi_0\) values. This means that if the same peptide is queried in two samples (run 1 with a low \(\pi_0\) value, and run 2 with a high \(\pi_0\) value) using the same parameters, and the scored peak groups have a similar discriminant score (\(d\)-score), then the peak groups might be assigned a low \(q\)-value in run 1 and a high \(q\)-value in run 2. Middle, the ‘experiment-wide context’ considers all runs of an experiment for error-rate control. The resulting \(q\)-values can be compared in terms of peak group quality between runs but should not be considered outside the context of the whole experiment. Bottom, the ‘global context’ only considers the best-scoring peak group per analyte across the entire experiment (as indicated by an asterisk). This approach enables the total set of detectable peptides or inferred proteins to be determined within the experiment. The global set of proteins can optionally be used as a constraint for the experiment-wide context to obtain the number of detected analytes in single runs.
high-scoring targets. The subsequent peptide-query-level q-value estimation step further uses the decoys to model a null distribution\(^2\). This concept can be extended to the protein level by applying a strategy similar to that suggested for discovery proteomics, by only considering the best-scoring PSM (or peptide query) for each peptide or protein for q-value estimation\(^22,28,29,48\).

We demonstrate the applicability of this extended q-value estimation approach on an exemplary sample (one run) of the SWATH-MS interlaboratory reproducibility study\(^37\) by querying for 194,052 proteotypic peptides contained in the CAL (Fig. 1 and Supplementary Figs. 1 and 2). Here, only the best-scoring peak group per protein (\(n = 10,316\)) was considered for protein-level q-value estimation. The discriminant score distributions and \(p\)-value histograms\(^43\) indicated that, in analogy to the peptide-query-level q-values, peptide-level and protein-level q-values can be applied as confidence metrics to avoid error accumulation from the peptide-query level to the protein level.

**Context-dependent estimation of error rates**

The q-value estimation for individual peptide queries (or proteins) is dependent on the context of the query, i.e., on other queries to the data\(^18\). This encompasses all peptide queries in the same LC-MS/MS run, and, in the context of a multi-sample study, also the peptide queries in different LC-MS/MS runs. In an individual run, one can ask whether the query peptide is detected in that particular sample\(^4\). If several runs are compared, then one could ask in which subset of samples the query peptide can be detected in. Alternatively, it might be of interest to see whether the query peptide can be detected in at least one of the samples of a study. Depending on which question needs to be answered, the context of the hypothesis and the method for estimating an appropriate q-value need to be adjusted. Analogously to the spectrum-centric approaches (Supplementary Note 4), we suggest considering three scenarios for DIA peptide-centric scoring and error-rate control: a run-specific context, an experiment-wide context and a global context.

**Run-specific context.** To investigate which peptides can be detected within one LC-MS/MS run (i.e., one sample injection), the run-specific context applies. The FDR or q-values are therefore estimated from the single best-scoring peak group per peptide query within one specific run, independently from those of the other runs that may have been acquired in the course of an experiment. Given a specified confidence threshold, the number of detectable peak groups, peptides or inferred proteins per run can be compared to the numbers achieved in other runs. This mode offers granularity for different levels of target peptide prevalence, as \(\pi_0\) values are estimated for each run separately. Samples with a low \(\pi_0\) value thus benefit in sensitivity, because only a limited multiple-testing correction is required. In contrast, samples with a high \(\pi_0\) value are more strictly corrected for multiple-testing\(^43\). This effect has various implications for the analysis of comparative studies containing heterogeneous samples with truly different \(\pi_0\) values between runs, for example, in affinity purification–mass spectrometry (AP-MS) experiments or fractionated samples. This means that if peptides are queried using parameters based on the same spectral library against two runs that result in the same \(\pi_0\) value, then peak groups with identical discriminant scores will also have the same estimated q-values. However, if the same peptides are queried against two runs with truly different \(\pi_0\) values, then peak groups with identical discriminant scores will have substantially different q-values (Fig. 2).

**Experiment-wide context.** This scenario is used to investigate in which subset of samples the query peptide is detected. In contrast to the run-specific context, the experiment-wide context assesses detected peptides and inferred proteins within an experiment that consists of multiple runs, and it estimates the \(\pi_0\) value from the best-scoring peak group matrix over all of the peptide queries and runs. A main assumption of this type of analysis is that the \(\pi_0\) values between individual runs are different, because the samples represent different proteome subsets (for example, comparison of a whole-cell lysate and fractionated samples) but not because of varying quality of the samples or runs. These conditions are more frequently met in peptide-centric than in spectrum-centric scoring methods, because in comparative studies individual runs are queried for many peptides that might not be detectable in the sample. In this case, peptides with identical q-values will have an identical discriminant score (Fig. 2).

Both the run-specific and experiment-wide contexts can be used to generate matrices of detectable peak groups, peptides or inferred proteins while controlling the error rate. However, when the analytes are summarized across a large study, false-positive
detections from individual samples are accumulated. This effect is illustrated in Figure 3, which shows the cumulatively detected peak groups and inferred proteins across the 229 runs that constitute the interlaboratory SWATH-MS study37 and, independently, across the 246 runs of a previously published study that measured undepleted human blood plasma samples from 116 individuals38. The corresponding decoy accumulation rate is shown in Supplementary Figures 3–6. When using the CAL and applying a q-value cutoff of 1% on the peptide-query level, as estimated within the experiment-wide context, the cumulative number of target proteins inferred reached almost the number of proteins covered by the spectral library (Fig. 3c). Applying an experiment-wide context with a q-value cutoff of 1% on the peptide-query level and a q-value cutoff of 1% on the protein level decreased the number of inferred proteins but still resulted in an accumulation of detected peptides and inferred proteins in the HEK-293 samples (Fig. 3a,c). This was not the case for the samples of the plasma data set, which contained more peptides per inferred protein on average, but a much lower total number of proteins (Fig. 3b,d). To prevent such an accumulation of potentially false-positively analytic detections in affected studies, the global context can be applied.

Global context. This context is used to investigate which peptides can be detected in at least one LC-MS/MS run of the experiment. For this purpose, it considers only the best-scoring detected peak groups, peptides or inferred proteins over all of the runs in a study for the error-rate control. The resulting global protein master list is a set of overall inferred proteins in the entire study that can be used to filter the matrix obtained by using either the run-specific or the experiment-wide context. The effect of applying constraints based on the global context is shown in Figure 3. The use of a peptide-query-level and protein-level global FDR cutoff of 1% (in addition to the 1% peptide-query-level, experiment-wide FDR cutoff) resulted in a consistent number of cumulatively detected analytes across all 229 runs of the interlaboratory SWATH-MS study, even when we used the large CAL (Fig. 3). In the plasma data set, accumulation at the inferred protein level was already reasonably well controlled by the experiment-wide FDR at the protein level, and the application of the global context constraint further reduced the observed accumulation of false-positively inferred proteins (Fig. 3b,d).

Tradeoff between spectral library specificity and comprehensiveness
As discussed above, sample-specific spectral libraries have the benefit of less error-rate control being required (low π0 value); however, the achievable proteome coverage depends on the completeness of the library. In contrast, repository-scale spectral libraries, which cover additional peptides that are detectable in the sample but are not found in the sample-specific spectral library, can reach higher coverage of the studied proteome when additionally detected peptides are not lost to the stricter multiple-testing adjustments (high π0 value) required. Adding new undetectable targets will only reduce sensitivity when the multiple-testing correction is appropriately applied, as was demonstrated in a recent study39 (Supplementary Note 3).

To further illustrate these effects, we applied here three spectral libraries, each with different levels of sample specificity and comprehensiveness, to query the data of the interlaboratory SWATH-MS study37. A sample-specific library (SSL) was generated from the spectra obtained in six DDA runs of the SWATH-MS interlaboratory study sample. The CAL was used as a second, repository-scale library, which consists of 331 runs of which 134 were acquired from fractionated and unfractionated HEK-293 samples. The third library applied was a HEK-293 subset of the CAL (HEK) and a sample-specific spectral library (SSL). (b) Discriminant score density plots on the protein level for queries in the global context with the CAL, HEK and SSL. (c) Comparison of the sets of proteins inferred at a 1% protein FDR in the global context of all 229 DIA runs of the SWATH-MS interlaboratory comparison study using the CAL, HEK and SSL spectral libraries. (d–f) Venn diagrams (d,f) and discriminant score density plots (e) showing the corresponding results of the analysis on the plasma data set.

Figure 4a illustrates the size and protein overlap between the different libraries used for the analysis of the interlaboratory SWATH-MS study. In Figure 4b, the global protein-level discriminant score distributions of targets and decoys are shown, which illustrates the different π0 values between the libraries. The reported proteins were compared after independent q-value estimation at a global protein-level cutoff of 1% (Fig. 4c). When applying peptide queries based on the HEK-293 sample-specific spectral library, all of the proteins could be recovered from the DIA data of the interlaboratory SWATH-MS study. For the queries based on the CAL, a global set of 4,989 proteins was inferred at a 1% protein FDR. This corresponded to a protein-level recovery of ~50% relative to the number of proteins in the CAL and was almost twice the number of proteins that could be inferred by using the sample-specific spectral library, indicating that the additional proteins were not identified in the sample-specific DDA runs or did not fulfill the requirements for peptide-query parameter
generation. For the HEK-293 subset of the CAL, 4,841 of the 6,019 proteins queried were confidently inferred. The relatively small discrepancy between the proteins inferred using the CAL and its HEK-293 subset illustrates the tradeoff of a larger, but more comprehensive query space, requiring strict multiple-testing correction. The 380 (7.8% of total) proteins that were exclusively detected with the HEK-293 subset library illustrates a loss of sensitivity, whereas the additional 503 (10% of total) proteins detected with the CAL illustrates the opportunity gained.

Figure 4d–f illustrates the size and protein overlap between the different libraries, the global protein-level discriminant score distributions, and the reported protein overlap at 1% global protein-level FDR, respectively, for the plasma data set. Even though the subset of proteins that can be inferred in the plasma data set is smaller than the ones from the interlaboratory SWATH-MS study, the relative results are qualitatively similar. On the one hand, this analysis showed that large comprehensive spectral libraries can achieve sensitive results at an appropriate error-rate control. On the other hand, decreasing the number of peptide queries can lower the requirement for multiple-testing adjustments at the potential cost of proteome coverage. Therefore, the optimal trade-off for a study depends on how well the spectral library represents the actual sample content.

**DISCUSSION**

With the increasing numbers of peptides queried in samples acquired in data-independent acquisition mode by peptide-centric targeted data extraction, it is imperative to adopt strict quality assessment metrics, such as the established criteria from spectrum-centric discovery proteomics, to ensure reproducible reporting of results. Here we have discussed the challenges associated with error-rate control in the analysis of DIA data. We demonstrated that the FDR should be controlled not only on the peptide-query level, but also on the peptide and protein levels in peptide-centric scoring workflows that apply comprehensive spectral libraries. Furthermore, we propose the application of different context-dependent error-rate estimation strategies. While the run-specific context offers per-run granularity, the experiment-wide context provides comparable result matrices across large heterogeneous data sets. The global context can be used to generate a list of detected peak groups, peptides and inferred proteins that can be confidently detected in a study. We suggest that a practical method to control the error rate is to filter the result matrices generated from either the run-specific or experiment-wide contexts using the set of analytes that are confidently detected in the global context. We have shown that this results in a uniform set of inferred proteins with negligible accumulation of false positives over a large number of samples.

The error-rate control strategies we have described are implemented and available in an updated PyProphet version (Online Methods) and are available in Spectronaut 11 (ref. 13). Future developments might extend the statistical models to adjust probabilities for the detection of peptides and inference of proteins across multiple runs to improve detection sensitivity.26,47 Other extensions and adaptations may be necessary, for example, for heterogeneous data sets that are acquired on different instrument types and are analyzed together, or if the parameters and assumptions of the algorithms are changed (Supplementary Fig. 7 and Supplementary Note 5). Despite our proposed strategies to control error rates in large-scale targeted proteomics experiments, the increased query space in repository-scale spectral libraries compromises the detection sensitivity. We have illustrated with the interlaboratory SWATH-MS study37 and the plasma data set48 that different spectral library specificities and comprehensiveness have profound effects on the importance of multiple-testing corrections and the respective results of the analyses. Therefore, it might be interesting for future applications to consider strategies for reducing the query space to provide an optimal tradeoff between proteome coverage and the fraction of undetectable targets. For this purpose, several different strategies have been suggested previously (Supplementary Note 6); however, further investigations are required to evaluate the optimal tradeoffs for different studies. The future development of algorithms will continue to abolish the borders between spectrum-centric and peptide-centric scoring approaches to provide fully integrated workflows.

The development and application of DIA as an enabling tool in quantitative proteomics has undergone rapid expansion in recent years, and this is set to continue for the foreseeable future. We hope to stimulate community discussion on these topics, and to aid researchers in choosing appropriate strategies for error-rate control to broadly improve the quality of data emerging from DIA-based quantitative proteomics studies.

**METHODS**

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

**Note:** Any Supplementary Information and Source Data files are available in the online version of the paper.

**ACKNOWLEDGMENTS**

Please note that M.H., C.L.H., Y.L., M.J.M., B.X.M., A.I.N., P.G.A.P., L.R., H.L.R., S.T. and Y.S.T. were added to the author list in alphabetical order. We thank the authors of the SWATH-MS interlaboratory study and of the human blood plasma data set for providing the data to conduct this study. We also thank the Scientific IT Support (ID SIS) and the high-performance computing (HPC) teams of ETH Zurich for support and maintenance of the computing infrastructure. M.H. was supported by a grant from the Institut Mérieux; A.I.N. was funded by the US National Institutes of Health (NIH; grant R01GM094231); H.L.R. was funded by the Swiss National Science Foundation (SNSF; grant P2ZEP3 162268); B.C.C. was supported by a SNSF Ambizione grant (PZ00P3_161435); and R.A. was supported by ERC Proteomics v3.0 (Adg-233226 Proteomics v.3.0) and Adg-670821 Proteomics 4D, the PhosphonetX project of SystemsX.ch and the Swiss National Science Foundation (SNSF) grant 31003A_166415.

**AUTHOR CONTRIBUTIONS**

G.R., I.B. and R.A. wrote the paper with feedback from all authors; G.R. and B.C.C. developed the methods; I.B. analyzed the data set; U.S. and G.R. implemented the PyProphet extension; M.H., C.L.H., Y.L., M.J.M., B.X.M., A.I.N., P.G.A.P., L.R., H.L.R., S.T. and Y.S.T. provided critical input on the project; and B.C.C. and R.A. designed and supervised the study.

**COMPETING FINANCIAL INTERESTS**

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Domon, B. & Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. Nat. Biotechnol. 28, 710–721 (2010).

2. Chapman, J.D., Goodlett, D.R. & Masselton, C.D. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. Mass Spectrom. Rev. 33, 452–470 (2014).
3. Gillett, L.C., Leitner, A. & Aebersold, R. Mass spectrometry applied to bottom-up proteomics: entering the high-throughput era for hypothesis testing. *Annu. Rev. Anal. Chem.* (Palo Alto Calif.) 9, 449–472 (2016).
4. Ting, Y.S. et al. Peptide-centric proteome analysis: an alternative strategy for the analysis of tandem mass spectrometry data. *Mol. Cell. Proteomics* 14, 2301–2307 (2015).
5. Silva, J.C. et al. Quantitative proteomic analysis by accurate mass-retention-time pairs. *Anal. Chem.* 77, 2187–2200 (2005).
6. Tsou, C.-C. et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* 12, 258–264 (2015).
7. Wang, J. et al. MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. *Nat. Methods* 12, 1106–1108 (2015).
8. Li, Y. et al. Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. *Nat. Methods* 12, 1105–1106 (2015).
9. Gillett, L.C. et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* 11, 0111.016717 (2012).
10. Röst, H.L. et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* 32, 219–223 (2014).
11. Telemann, J. et al. DIANA—algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics* 31, 555–562 (2015).
12. MacLean, B. et al. Skyline: an open-source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 26, 966–968 (2010).
13. Bruderer, R. et al. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics* 14, 1400–1410 (2015).
14. Carr, S.A. et al. Targeted peptide measurements in biology and medicine: best practices for mass-spectrometry-based assay development using a fit-for-purpose approach. *Mol. Cell. Proteomics* 13, 907–917 (2014).
15. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Stat. Methodol.* 57, 289–300 (1995).
16. 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. *Anat. Chem.* 74, 5383–5392 (2002).
17. Choi, H. & Nesvizhskii, A.I. Semi-supervised model-based validation of peptide identifications in mass-spectrometry-based proteomics. *J. Proteome Res.* 7, 254–265 (2008).
18. 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.* 7, 40–44 (2008).
19. Genovese, C. & Wasserman, L. Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. B Stat. Methodol.* 64, 499–517 (2002).
20. Iyer, V. & Sarkar, S. An adaptive single-step FDR procedure with applications to DNA microarray analysis. *Biom. J.* 49, 127–135 (2007).
21. Storey, J.D. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* 31, 2013–2035 (2003).
22. Nesvizhskii, A.I. A survey of computational methods and error-rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* 73, 2092–2123 (2010).
23. Käll, L., Canterbury, J.D., Weston, J., Noble, W.S. & MacCoss, M.J. Semi-supervised learning for peptide identification from shotgun proteomics data sets. *Nat. Methods* 4, 923–925 (2007).
24. Serang, O. & Noble, W. A review of statistical methods for protein identification using tandem mass spectrometry. *Stat. Interface* 5, 3–20 (2012).
25. The, M., Tasnim, A. & Käll, L. How to talk about protein-level false discovery rates in shotgun proteomics. *Proteomics* 16, 2461–2469 (2016).
26. Shteynberg, D. et al. iProphet: multilevel integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* 10, M111.007690 (2011).

27. Reiter, L. et al. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* 8, 2405–2417 (2009).
28. Savitski, M.M., Wilhelm, M., Hahne, H., Kuster, B. & Bantscheff, M. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Mol. Cell. Proteomics* 14, 2394–2404 (2015).
29. The, M., MacCoss, M.J., Noble, W.S. & Käll, L. Fast and accurate protein false discovery rates on large-scale proteomics data sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.* 27, 1719–1727 (2016).
30. Choi, H., Ghosh, D. & Nesvizhskii, A.I. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J. Proteome Res.* 7, 286–292 (2008).
31. Ahrens, C.H., Brunner, E., Qeli, E., Basler, K. & Aebersold, R. Generating and navigating proteome maps using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* 11, 789–801 (2010).
32. Reiter, L. et al. mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods* 8, 430–435 (2011).
33. Karlsson, C., Malmström, L., Aebersold, R. & Malmström, J. Proteome-wide selected reaction monitoring assays for the human pathogen Streptococcus pyogenes. *Nat. Commun.* 3, 1301 (2012).
34. Schubert, D.T. et al. The MtB proteome library: a resource of assays to quantify the complete proteome of *Mycobacterium tuberculosis*. *Cell Host Microbe* 13, 602–612 (2013).
35. Picotti, P. et al. A complete mass spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* 494, 266–270 (2013).
36. Rosenberger, G. et al. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data* 1, 140031 (2014).
37. Collins, B.C. et al. Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH–mass spectrometry. *Nat. Commun.* 8, DOI: 10.1038/s41467-017-00249-5 (2017).
38. Liu, Y. et al. Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* 11, 766 (2015).
39. Seeleven, N. et al. Reproducible and consistent quantification of the *Saccharomyces cerevisiae* proteome by SWATH-MS. *Mol. Cell. Proteomics* 14, 739–749 (2015).
40. Guo, T. et al. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nat. Med.* 21, 407–413 (2015).
41. Schubert, D.T. et al. Absolute proteome composition and dynamics during dormancy and resuscitation of *Mycobacterium tuberculosis*. *Cell Host Microbe* 18, 96–108 (2015).
42. Schubert, D.T. et al. Building high-quality assay libraries for targeted analysis of SWATH-MS data. *Nat. Protoc.* 10, 426–441 (2015).
43. Storey, J.D. & Tibshirani, R. Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. USA* 100, 9440–9445 (2003).
44. Serang, O. & Käll, L. Solution to statistical challenges in proteomics is more statistics, not less. *J. Proteome Res.* 14, 4099–4103 (2015).
45. Blattmann, P., Heusel, M. & Aebersold, R. SWATH2stats: an R/Bioconductor package to process and convert quantitative SWATH-MS proteomics data for downstream analysis tools. *PloS One* 11, e0153160 (2016).
46. Tsou, C.-C., Tsai, C.F., Teo, G.C., Chen, Y.J. & Nesvizhskii, A.I. Untargeted, spectral library-free analysis of data-independent acquisition proteomics data generated using Orbitrap mass spectrometers. *Proteomics* 16, 2257–2271 (2016).
47. Keller, A., Bader, S.L., Shteynberg, D., Hood, L. & Mortiz, R.L. Automated validation of results and removal of fragment ion interferences in targeted analysis of data-independent acquisition mass spectrometry (MS) using SWATHProphet. *Mol. Cell. Proteomics* 14, 1411–1418 (2015).
48. Gupta, N. & Pevzner, P.A. False discovery rates of protein identifications: a strike against the two-peptide rule. *J. Proteome Res.* 8, 4173–4181 (2009).
49. Muntel, J. et al. Advancing urinary protein biomarker discovery by data-independent acquisition on a quadrupole-orbitrap mass spectrometer. *J. Proteome Res.* 14, 4752–4762 (2015).
We implemented the described error-rate control strategies in an updated and extended version of PyProphet\(^1\). PyProphet is a Python-based re-implementation of the mProphet\(^2\) algorithm that was originally developed for semi-supervised learning and the statistical validation of targeted proteomics data. The PyProphet implementation procedure reported here extended the original approach by the following options:

**Semi-supervised learning.** Instead of conducting independent iterations of learning and statistical validation separately per run, PyProphet conducts subsampling of paired target and decoy peak groups\(^3\) over all runs to learn a single, experiment-wide linear discriminant analysis (LDA) scoring model. From the LDA function, a discriminant score is derived by computation of the \(z\)-score by using the decoy peak group mean and s.d. as described previously\(^4\). The purpose of this integrated step is to ensure that the peak groups can be sorted according to their quality in a unified way across heterogeneous samples or samples of variable quality.

**Statistical validation.** In addition to the original parametric assumptions\(^5\), PyProphet now also supports nonparametric, empirical estimation of \(p\)-values\(^6\). To estimate \(q\)-values on different levels, PyProphet enabled aggregation over peptide-level or protein-level groups by selection of the best-scoring peak group. For each level, the \(q\)-values, the FDR\(^7\) and FNR\(^8\,9\) or the positive FDR (pFDR) and positive FNR (pFNR)\(^10\,11\) were computed independently using the corresponding decoys as the null model. For the different contexts, PyProphet supports different modes to either conduct \(q\)-value estimation per run (run-specific context), across all runs (experiment-wide context) or in a global fashion (global context).

**Multi-run and high-throughput processing.** To process large data sets, such as that in the interlaboratory SWATH-MS study\(^3\), we improved the scalability of PyProphet under conditions in which hundreds of runs, each with a file size of 5–10 GB, needed to be analyzed concurrently. The new PyProphet version was optimized for parallel processing in a cluster environment (IBM Platform LSF or OpenLava), but it can readily be adopted to other environments with Python extensions. By using subsampling and integrated scoring, \(q\)-value estimation can be conducted using very large numbers of peptide queries for hundreds of runs within hours by using a common cluster or cloud environment. A full analysis of the 229 OpenSWATH reports of individually analyzed runs (9 GB per run) using 1–32 central processing units (CPUs) (depending on the individual step) and 4–48 GB RAM (depending on the individual step) required a processing time of 1.5 h, using several sequential and parallel jobs. Because the OpenSWATH results were stored as text files, the main requirement for the processing was throughput of filesystem input/output operations and temporary storage capacity.

**Code availability.** Our software is implemented in Python, available for all major platforms and released under the 3-clause BSD license. PyProphet is available along with detailed instructions from https://github.com/PyProphet. Further documentation of our workflow is available on http://openswath.org.

Analysis of the SWATH-MS interlaboratory reproducibility data set. **Generation of the combined human assay library (CAL).** The CAL for the 64-variable-windows setting\(^1\) was filtered for proteotypic peptides and complemented by 30 additional stable-isotope-labeled standard (SIS) peptides as described previously\(^3\).

**Generation of the HEK-293 subset library.** The HEK-293 subset library (HEK) was generated by filtering the CAL to contain only spectra from HEK-293 samples. The peptide-query parameters were derived from the HEK-293-filtered spectral library as described previously\(^6\).

**Generation of the sample-specific library.** The sample-specific library (SSL) was generated from the spectra collected by six LC-MS/MS runs in DDA mode of the identical unfractonated HEK-293 tryptic digest, as described previously\(^6\).

**Generation of the combined human + Mycobacterium tuberculosis library.** Based on the SpectraST consensus library of the CAL\(^3\) and the M. tuberculosis\(^34\,35\) libraries, we generated a merged library by appending the M. tuberculosis library to the CAL library using SpectraST (TPP 5.0). The protein identifiers were updated using the combined original FASTA files of the two libraries, to later exclude any shared peptides between the two organisms. Peptides and fragment ions were selected identically as described previously\(^3\,36\) using msproteomicstools (Git version master@e10a2b88), and OpenMS (version 2.1) was used with OpenSwathDecoyGenerator to generate combined target-decoy libraries (method: shuffle; similarity_threshold: 0.05; identity_threshold: 0.7; exclude_similar: true; append: true).

**DIA data analysis.** The analysis of the interlaboratory SWATH-MS data set was conducted identically to that described in Collins et al.\(^37\). The analysis of the SWATH-MS data was performed using OpenSWATH (OpenMS v2.0), essentially as previously described\(^10\) but with the following modified parameters: \(m/z\) extraction window = 75 p.p.m., and retention time (RT) extraction window = 900 s. The analysis was performed separately for the four different spectral libraries described above: the CAL, the HEK-293 subset library, the sample-specific library and the combined human + M. tuberculosis library.

Semi-supervised learning and statistical validation were performed using the extended version of PyProphet described above (PyProphet-cli v0.19, https://github.com/PyProphet). PyProphet was run for all three available contexts to conduct \(q\)-value estimation per run (run-specific context), across all runs (experiment-wide context) or in a global fashion (global context), with a fixed \(\lambda\) of 0.4. The set of peptide peak groups used for learning the score weights of OpenSWATH subcores to produce a single discriminant score were sampled with a ratio \(=1/(\text{number of samples})\); for aggregated analysis of all sites, a ratio of 0.005 was used.

A global ‘master list’ of detected peak groups and proteins across the entire data set was generated by filtering the results from the global context at 1% peptide-query FDR and 1% protein FDR. The results from the experiment-wide context were filtered at three different stringency levels: 1% peptide-query FDR, 1% peptide-query FDR and 1% protein FDR, and 1% peptide-query FDR and additional filters based on the global master list of peptide queries and proteins.

For the analysis of the three different libraries (Fig. 4), separate scoring models were trained.

**Analysis of the plasma data set.** The CAL for the 32-fixed-windows setting\(^36\) was used to analyze the plasma data set\(^38\) similarly to that described for the interlaboratory SWATH-MS study data set described above, with the following differences for OpenSWATH:
m/z extraction window = 0.05 Da, and RT extraction window = 600 s. The following set of scores was used: xx_lda_prelim_score, intensity_score, isotope_correlation_score, isotope_overlap_score, library_corr, library_rmsd, log_sn_score, massdev_score, massdev_score_weighted, norm_rt_score, xcorr_coelution, xcorr_coelution_weighted, courtship, and xcorr_shape_weighted.

The OpenSWATH results were filtered to contain only proteotypic peptides. To generate the results for the plasma subset analysis, the OpenSWATH results were filtered to contain only peptides mapping to proteins that were confidently detected (confidence threshold for inclusion in original library 36) in at least one of the eight DDA plasma runs that were part of the CAL. This approach was equivalent to using a subset library for data extraction by OpenSWATH.

PyProphet was executed as described above; however, the scoring model (LDA weights) of the plasma subset analysis was applied to the whole CAL analysis to ensure that the differences of the comparison originated only from the different library sizes.

The analysis was conducted independently for both the parametric and the nonparametric methods.

**Data availability.** The raw data and processed results of the analysis of the SWATH-MS interlaboratory reproducibility study have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository 50 with the data set identifier PXD004884.

The processed results of the analysis of the plasma data set have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository 50 with the data set identifier PXD006625. Source data for Figures 1, 3 and 4 are available online.

A Life Sciences Reporting Summary is available.

50. Vizcaíno, J.A. et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* 41, D1063–D1069 (2013).
Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

▶ Experimental design

1. Sample size
   Describe how sample size was determined.
   This study focused on the reanalysis of already published datasets: (1) SWATH-interlaboratory study (Collins et al. 2017) with 229 samples and (2) human plasma study (Liu et al. 2015) with 246 samples.

2. Data exclusions
   Describe any data exclusions.
   All data from the previously published studies were considered.

3. Replication
   Describe whether the experimental findings were reliably reproduced.
   Reproducibility of the described approach is demonstrated based on similar findings for the two independent studies: (1) SWATH-interlaboratory study (Collins et al. 2017) and (2) human plasma study (Liu et al. 2015).

4. Randomization
   Describe how samples/organisms/participants were allocated into experimental groups.
   See per original report: (1) SWATH-interlaboratory study (Collins et al. 2017) and (2) human plasma study (Liu et al. 2015).

5. Blinding
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   See per original report: (1) SWATH-interlaboratory study (Collins et al. 2017) and (2) human plasma study (Liu et al. 2015).

   Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters
   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

   n/a Confirmed

   □ □ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)

   □ □ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

   □ □ A statement indicating how many times each experiment was replicated

   □ □ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)

   □ □ A description of any assumptions or corrections, such as an adjustment for multiple comparisons

   □ □ The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted

   □ □ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)

   □ □ Clearly defined error bars

   See the web collection on statistics for biologists for further resources and guidance.

Nature Methods: doi:10.1038/nmeth.4398
Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

Our software PyProphet is available along with detailed instructions on https://github.com/PyProphet. All other software including descriptions are covered here http://www.openswath.org

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). Nature Methods guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

All data and software are freely available.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

n/a

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

As original report on SWATH-interlaboratory study (Collins et al. 2017)

b. Describe the method of cell line authentication used.

As original report on SWATH-interlaboratory study (Collins et al. 2017)

c. Report whether the cell lines were tested for mycoplasma contamination.

As original report on SWATH-interlaboratory study (Collins et al. 2017)

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

As original report on SWATH-interlaboratory study (Collins et al. 2017)

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

n/a

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

As original report on human plasma study (Liu et al. 2015)