Statistical control of peptide and protein error rates in large-scale targeted DIA analyses: Supplementary Notes

George Rosenberger[1,2,*], Isabell Bludau[1,2,*], Uwe Schmitt[3], Moritz Heusel[1,4,§], Christie Hunter[5,§], Yansheng Liu[1,§], Michael J. MacCoss[6,§], Brendan X. MacLean[6,§], Alexey I. Nesvizhskii[7,8,§], Patrick G. A. Pedrioli[1,§], Lukas Reiter[9,§], Hannes L. Röst[1,§], Stephen Tate[10,§], Ying S. Ting[6,§], Ben C. Collins[1,‡], Ruedi Aebersold[1,11,‡]

1 Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, CH-8093 Zurich, Switzerland.

2 PhD Program in Systems Biology, University of Zurich and ETH Zurich, CH-8093 Zurich, Switzerland.

3 ID Scientific IT Services, ETH Zurich, CH-8092 Zurich, Switzerland.

4 PhD program in Molecular and Translational Biomedicine, Competence Center Personalized Medicine (CC-PM), ETH Zurich and University of Zurich, CH-8044 Zurich, Switzerland.

5 SCIEX, 1201 Radio Road, Redwood City, CA 94065, USA.

6 Department of Genome Sciences, University of Washington, Seattle, WA 98195–5065, USA.

7 Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA.

8 Department of Pathology, University of Michigan, Ann Arbor, MI 48109, USA.

9 Biognosys, Wagistrasse 25, CH-8952 Schlieren, Switzerland.

10 SCIEX, Concord, Ontario L4K 4V8, Canada

11 Faculty of Science, University of Zurich, CH-8057 Zurich, Switzerland.

* Equal contribution

§ Authors ordered alphabetically

‡ Corresponding authors: Correspondence to aebersold@imsb.biol.ethz.ch or collins@imsb.biol.ethz.ch

Supplementary Notes

1. False non-discovery rate

Control of the FNR is not yet conducted routinely in proteomics, potentially because interpretation of the results and most downstream analysis strategies are more affected by accumulation of false positives than accumulation of false negatives. To investigate whether the FNR could provide a useful additional metric for error rate control, we have implemented support for FNR and pFNR estimation in PyProphet. Applied to the SWATH-MS inter-laboratory reproducibility study, we estimated the FNR values for the global context on the protein level at 1% FDR to be 15% for the CAL, 25% for the HEK and 0% for the SSL library. Unfortunately, these values are challenging to interpret, because they must be evaluated in the context of the library that was used to query the data. Since the absolute number of true negatives is smaller for the HEK library compared to the CAL, the lower number of false negatives results in a large FNR. In the peptide-centric scoring approach, the FNR value thus directly depends on the completeness and specificity of the applied library and is difficult to compare between different spectral libraries. For example, the 0% FNR obtained for the SSL means that all proteins represented in the library are present in the data, but this does not provide any information about what part of the proteome was not covered by the peptide queries. The FNR is a useful statistic also for applications in proteomics, but applied to DIA data, we believe it is better suited for spectrum-centric or hybrid strategies and comparisons of related samples, analyzed using the same spectral library.

2. Comparison of spectrum-centric search and peptide-centric query space

The number of queries influences π_0 in both spectrum-centric as well as peptide-centric scoring approaches. However, one important distinction between the spectrum-centric and peptide-centric scoring approach is the definition and effect of the search or query space. In spectrum-centric analyses, the protein sequence database defines the search space and directly influences π_0 and thus the error rate estimation. In contrast, the query space in peptide-centric scoring is defined by the acquired DIA data[1] and can be refined by the peptide ion specific parameters such as the relative retention time window. Here, the query

space limits the number of detected peak groups competing for the best scoring evidence of detection without directly influencing π_0.

3. Tradeoff between spectral library specificity and comprehensiveness

In many studies, direct DDA analysis of the samples of interest resulted in a spectral library that is smaller than the set of detectable peptides in the corresponding DIA data, an effect that may have some instrument dependence. This is demonstrated by the observation that employing spectral libraries made by fractionation of the sample for DDA analysis, or the use of repository-scale spectral libraries can increase the sensitivity in the analysis of DIA data[2-4] at the cost of increasing π_0. Recent algorithmic developments for spectrum-centric analysis of DIA data like DIA-Umpire[5] support peptide queries based on sample-specific spectral libraries generated directly from the DIA data. However, since many proteomic studies focus on comparing different experimental conditions or perturbations, it is desirable to target peptides of interest across all samples. In this scenario, the peptide queries and correspondingly the individual π_0 will grow rapidly with the sample heterogeneity and cohort size. This is particularly relevant in clinical cohort studies, where a large number of related but different samples are compared.

In a recent study[6], Muntel *et al.* investigated the human urinary proteome in triplicate, using sample-specific libraries as well as the same combined human assay library (CAL)[3] used in this study. The data used to generate the CAL did not contain any measurements from urinary samples and therefore only partially represented the proteins and peptides contained in the urinary sample-specific libraries. The authors analyzed the triplicate samples using the respective spectral libraries independently to define a cumulative, total set of detectable peptides across each triplicate analysis. To assess the reproducibility of detection, they computed the fraction of peptides that could be detected in all replicates as comparison metric. For sample-specific spectral libraries, signals representing 69% of the globally detectable peptides were detected in all three replicates. In contrast, if the same DIA data were queried with the peptide query parameters of the complete CAL, only 26% of the globally detectable peptides were detected in all three replicates. To investigate the reason for this discrepancy, the authors generated a specific instance of a human spectral

library that consisted of the subset of peptides from the CAL that were also contained in the sample-specific spectral library. This resulted in consistent detection in all replicates for 71% of the globally detectable peptides, a value comparable to the results obtained from a sample-specific library. This indicates that the peptide query parameters derived from the CAL are similarly specific and sensitive as the ones from the sample-specific library. The authors concluded that due to a very high π_0, the observed effects on reproducibility originated from the multiple hypothesis testing correction. Especially in their specific situation, where the combined library covered the urinary proteome poorly, sample-specific spectral libraries perform superior. Therefore, in specific samples, where the peptide prevalence in the reference spectral library is likely to be low, such as urinary or plasma proteomes, AP-MS digests or other specific sub-proteomes, it is crucial to either optimize the library or to adjust the error rate controlling efforts.

4. Implementations for context-dependent estimation of error rates

Q-value or FDR estimation in different contexts has been implemented in several variations and under different names. For example, PeptideProphet[7], ProteinProphet[8], OpenSWATH[2] and Spectronaut[9] among many other algorithms provide metrics on a run-specific level. Percolator[10], mQuest/mProphet[11], iProphet[12] or TRIC[13] estimate the statistics in an experiment-wide context when applied to several runs together. Algorithms like iProphet use estimated posterior error probabilities that were individually computed per run with different π_0 (and optionally updated with evidence from other runs) to then estimate an experiment-wide FDR on peptide sequence-level. Mayu[14], Andromeda[15] and ProteinInferencer[16] are examples for tools that can provide statistics in a global context.

5. Instrument and algorithm-specific considerations

It is important to consider that the discussed effects of spectral library specificity and experimental contexts are valid only under the assumptions that the individual runs were acquired on similar instruments and analyzed with identical parameters. If these assumptions are not fulfilled, grouping per condition, e.g. per instrument or parameter set and separate error rate control is necessary[17]. Further, different computational methods and parameters might have an effect on the scale of error accumulation. For example, in

this comparison, we applied the original parametric model of mProphet[11]. However, non-parametric approaches[10,18] can be more appropriate if the parametric assumptions are not fulfilled. Applied to the plasma data set, we found that non-parametric approaches are much more restrictive, lowering the accumulation of peptide detections across the more than 200 runs (Supplementary Figure 7). Nevertheless, the error accumulates on the protein-level, illustrating that even improved scoring functions and confidence estimates require reporting of results at appropriate levels.

6. Strategies to reduce the query space for spectral libraries

We have previously described and implemented simple methods to filter peptide queries for protein identifier sets, generated according to specific research questions, e.g. preliminary candidates or disease association[3]. Other strategies based on prior knowledge, such as matched transcript data, could also facilitate a reduction in the number of queries. In spectrum-centric data analysis, reduction of the query space based on prior knowledge such as likelihood of observing a particular peptide or protein based on global GPMDB data[19] or using complementary data such as RNA-Seq[20], or using the knowledge derived from the data being analyzed, as in e.g. using iterative database searching[21] (reviewed e.g. in [22,23]), have proven to be useful strategies in specific applications[24]. However, in many studies, the proteins of interest are not known *a priori*. A recent publication adapting spectral library searching for DIA data suggested peptide query optimization directly from the data to decrease the number of absent peptides queried by peptide-centric targeted data extraction tools[25]. As with the strategies developed for spectrum-centric DDA data, data-driven reduction of putative not detectable targets in peptide-centric scoring is conceptually attractive because no prior knowledge would be required.

Supplementary Figures

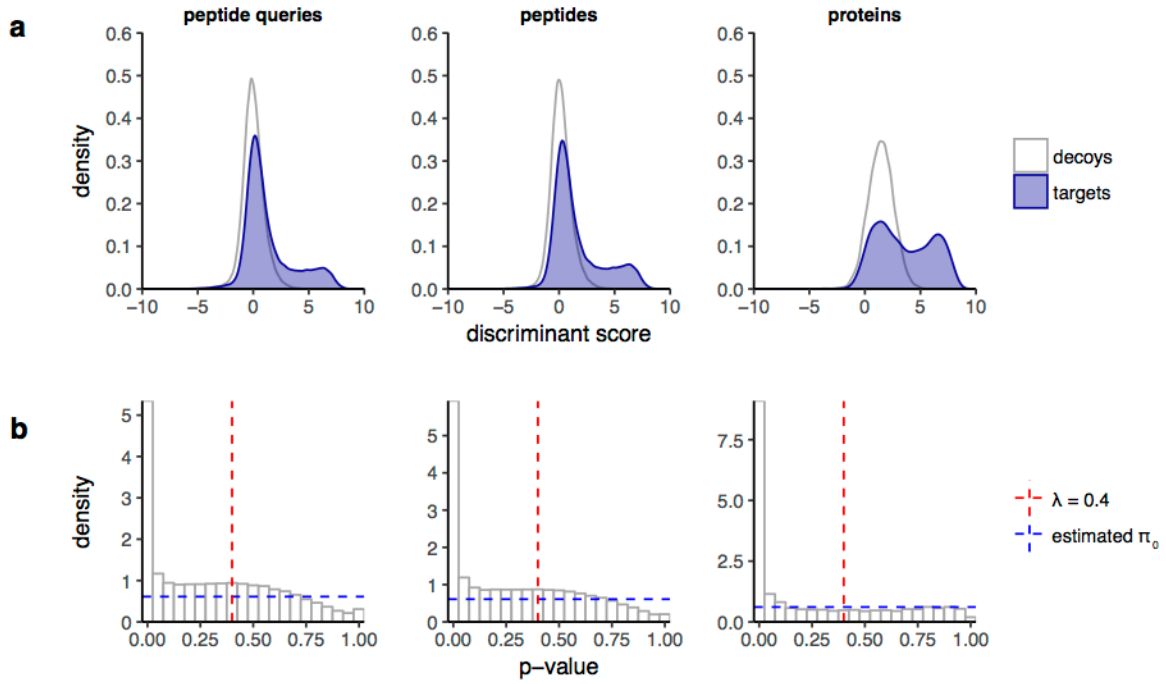

Figure S1. **Q-value estimation on peptide query-, peptide- and protein-level.** The peptide query-, peptide- and protein-level discriminant score density plots (**a**) and p-value histograms[26] (**b**) for one DIA run of the SWATH-MS inter-laboratory study analyzed with the combined human assay library (CAL) are depicted. **a)** The distributions indicate a large false target to total target ratio ($\pi_0 \approx 0.6$) on peptide query-level. The q-value estimation was adapted for peptide- and protein-level by using the best scoring peak group per peptide or protein across all samples for both targets and decoys. The false target to total target ratio decreases slightly on peptide-level and more on protein-level ($\pi_0 \approx 0.5$), compared to the peptide query-level. **b)** On peptide query- and peptide-levels, the estimation of π_0 is anticonservative, indicated by lower density of p-values after the p-value threshold of $\lambda = 0.4$. On the protein-level, the estimation of π_0 is more accurate with a consistent density of p-values[26].

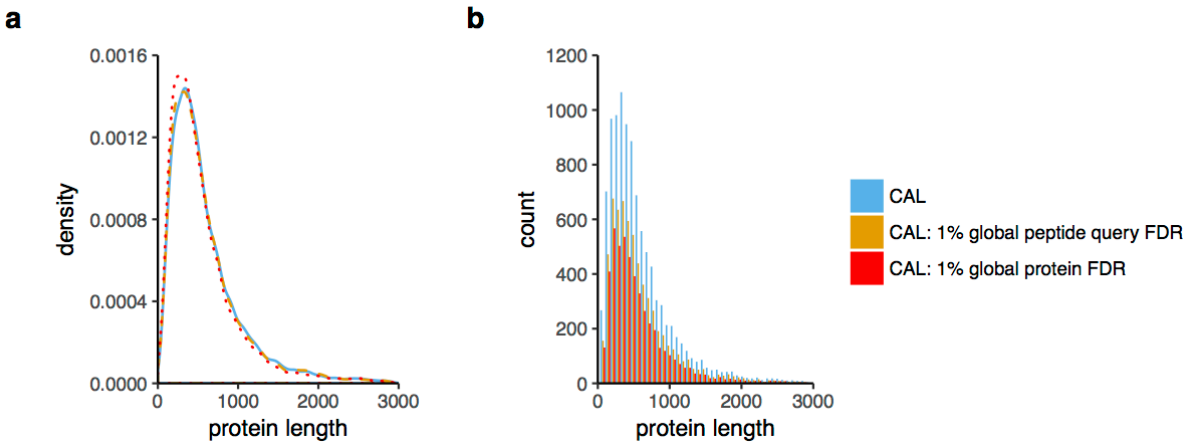

Figure S2. **Influence of protein length on the peptide query- and protein-level q-value estimation. a)** Protein length distribution of all proteins in the combined human assay library (CAL), all proteins inferred at 1% peptide query-level FDR in the global context of all 229 DIA runs of the SWATH-MS interlaboratory comparison study, and all proteins inferred at 1% global protein FDR respectively. **b)** Histogram of protein length distribution for the differently filtered protein subsets of the CAL. The distributions show that there is no bias for protein length when selecting the best peak group as proxy for protein-level q-value estimation.

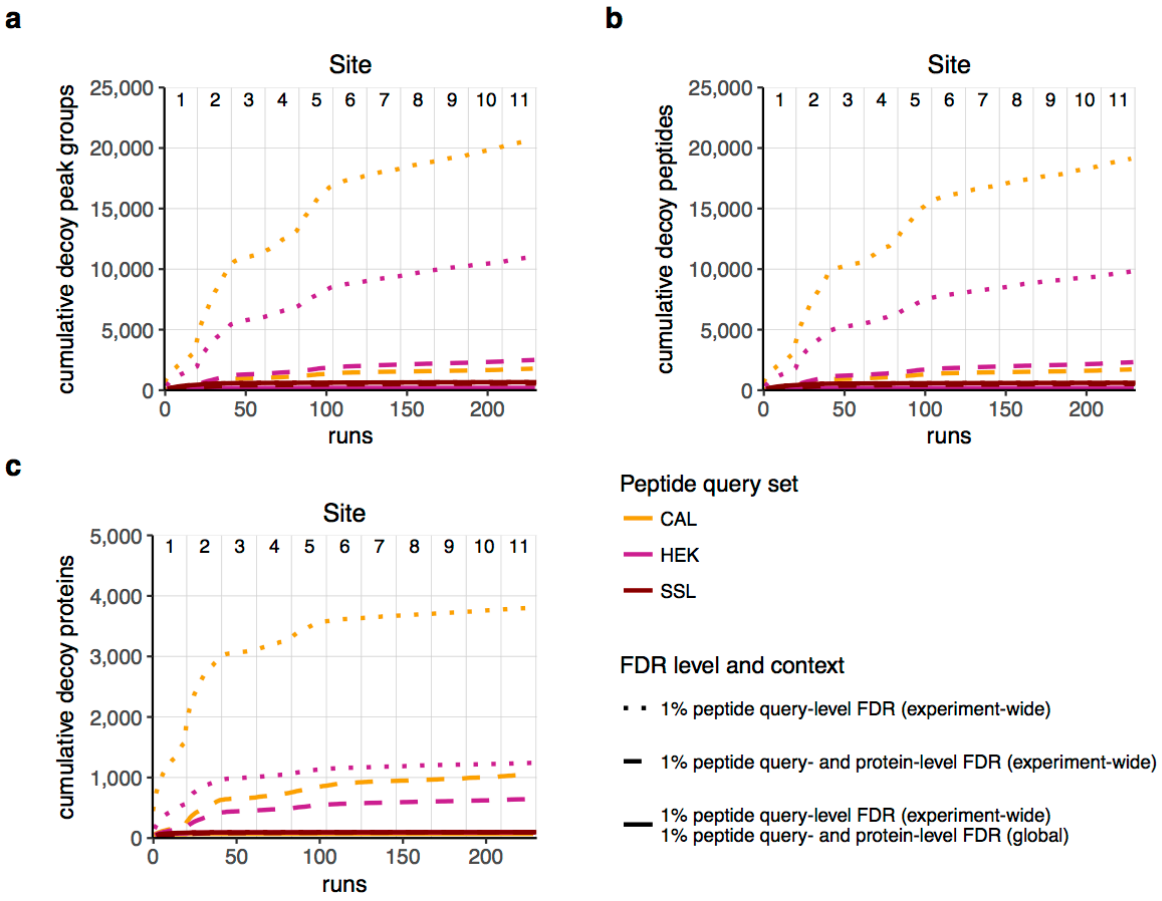

Figure S3. **Decoy accumulation across multiple runs.** The number of cumulatively detected peak group decoys (**a**), peptide decoys (**b**) and protein decoys (**c**) is shown for 229 DIA runs of the SWATH-MS inter-laboratory comparison data set[27].

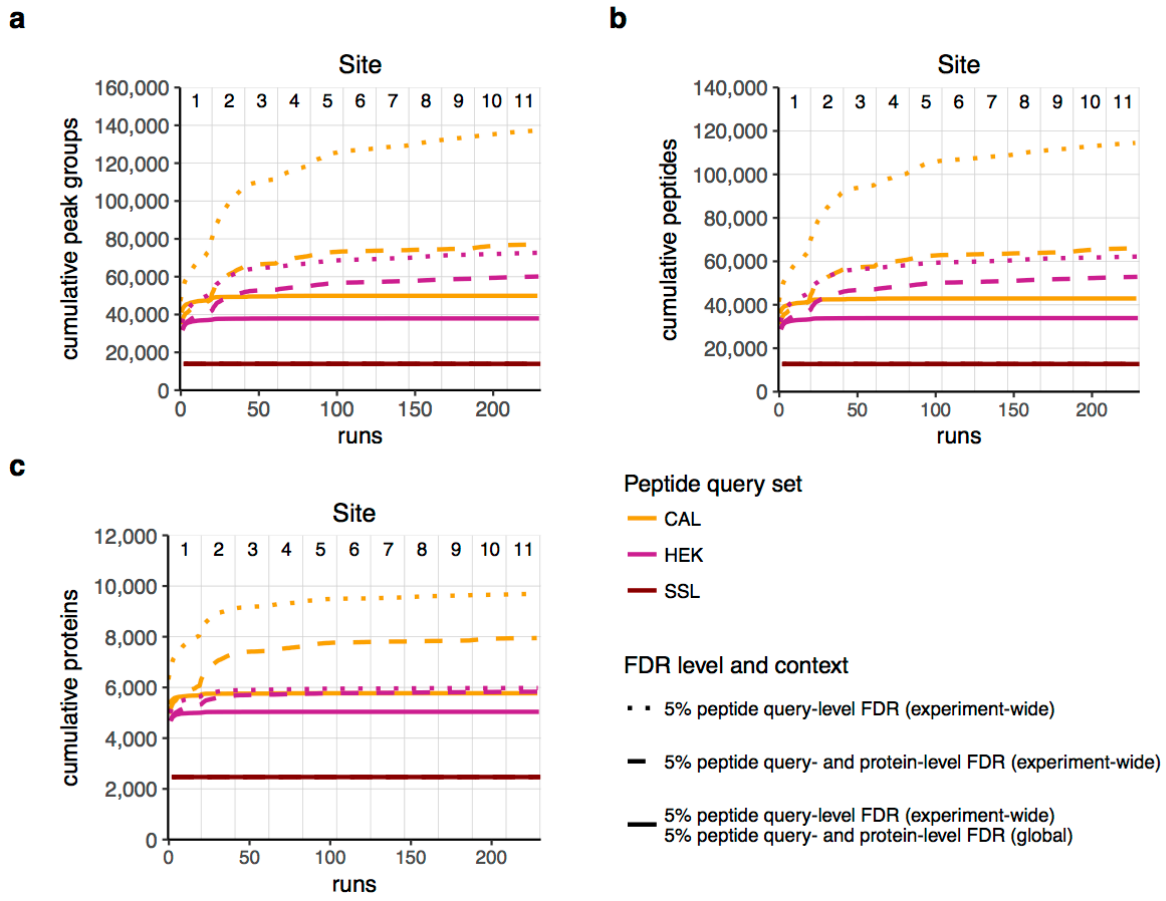

Figure S4. **Analyte accumulation across multiple runs (5% FDR).** The number of cumulatively detected peak groups (**a**), peptides (**b**) and proteins (**c**) is shown for 229 DIA runs of the SWATH-MS inter-laboratory comparison data set[27].

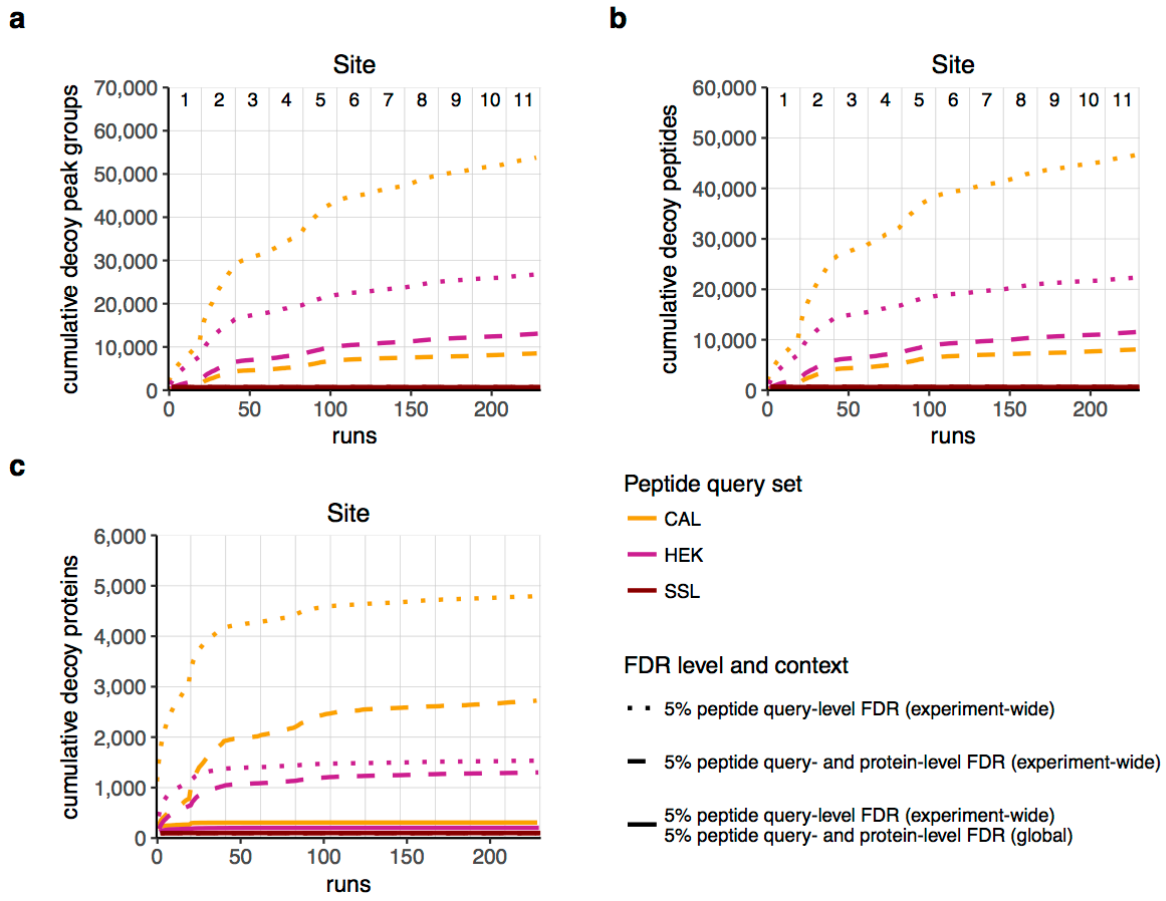

Figure S5. **Decoy accumulation across multiple runs (5% FDR).** The number of cumulatively detected peak group decoys (**a**), peptide decoys (**b**) and protein decoys (**c**) is shown for 229 DIA runs of the SWATH-MS inter-laboratory comparison data set[27].

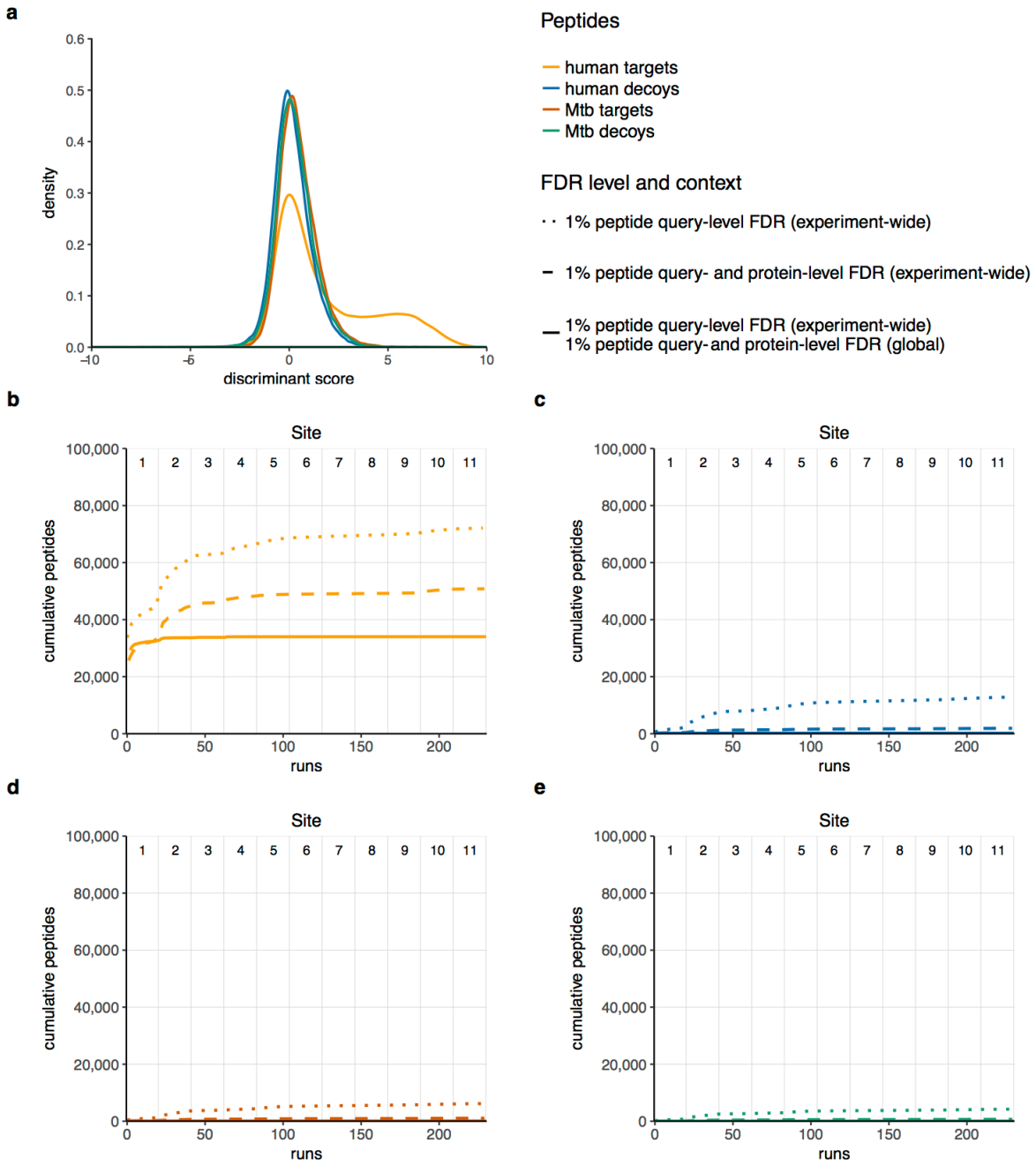

Figure S6. **Combined human and M. tuberculosis spectral library analysis. a**) The peptide-level discriminant score density of human targets, human decoys, *M. tuberculosis* (Mtb) targets, and Mtb decoys is shown for global analysis of the 229 DIA runs of the SWATH-MS inter-laboratory comparison data set[27] applying the combined human and Mtb spectral library. The Mtb targets and decoys show a similar distribution compared to the human decoys and the fraction of false human targets. The number of cumulatively detected peptides is shown for human targets (**b**), human decoys (**c**), Mtb targets (**d**), and Mtb decoys (**e**) from the combined human and Mtb spectral library with different error rate control

strategies. The Mtb decoy to target ratio is 0.82, explaining the absolute higher number of the accumulated Mtb targets.

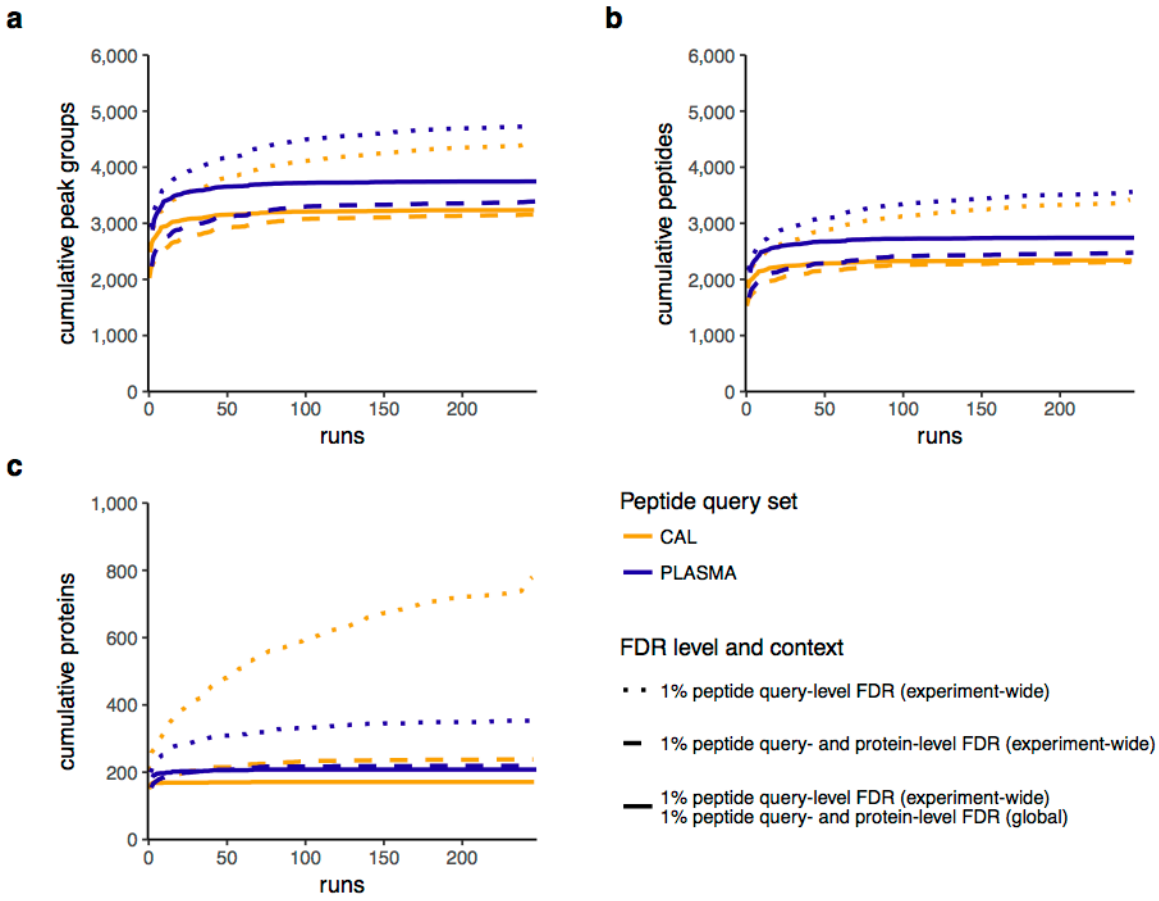

Figure S7. **Analyte accumulation across multiple runs in the plasma dataset (1% FDR).** The number of cumulatively detected peak groups (**a**), peptides (**b**) and proteins (**c**) is shown for the 246 DIA runs of the plasma data set[28] analyzed with the non-parametric model for q-value estimation.

Supplementary Tables

Term	*Definition*
Spectrum-centric scoring	Analysis of MS/MS spectra by assigning candidate peptide sequences from databases to each spectrum. "To which peptide sequence does the spectrum match best?"[1]
Peptide-centric scoring	Analysis of MS/MS spectra or ion chromatograms by querying peptides to detect evidence for analyte presence. "Is this peptide detected in the data?"[1]
Transition	Pair of precursor and product ion m/z.
Peptide query parameters	Empirically optimized set of transitions which in combination enable selective and sensitive detection of a peptide by a "peak group", co-eluting fragment ion chromatograms. Peptide query parameters (also referred to as "Tier 3" assays[29]) can be made more specific by including a normalized retention time[30] or empirical relative fragment ion intensities. Multiple sets of peptide query parameters, e.g. for different precursor charge states, can be used per peptide.
Peptide query	Targeted data extraction and scoring of DIA spectra using peptide query parameters resulting in candidate peak groups, of which commonly the best scoring is considered to originate from the target peptide.
Discriminant score	The discriminant score (also abbreviated as d-score here) is a combined score for each extracted peak group. It is computed in the semi-supervised learning step in PyProphet, from the initial peak group scores such as peak shape, co-elution, signal-to-noise-ratio, etc.
FDR	False discovery rate[31]; metric used for the control of the error rate of detected analytes in experiments affected by the multiple testing problem.
Q-value	Measure of significance of a detected analyte similar to the p-value, but accounting for the multiple testing problem analogously to the FDR.[26,32]
π_0	The prior probability that the null hypothesis is true, i.e. the ratio between undetectable targets (analytes that are not detectable in the queried sample) and the total number of queried targets.[26]
Analyte level	Peptide query-, peptide- or protein-level metrics, e.g. a protein q-value, representing the significance of an inferred protein.
Error rate context	Run-specific, experiment-wide or global contexts are used as attributes to specify whether an analyte metric, e.g. a protein q-value, should be interpreted independently per run, within the context of an experiment expression matrix or in a cumulative, global context.

Table 1. **Glossary and definitions.**

References

1. 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).
2. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32,** 219–223 (2014).
3. Rosenberger, G. *et al.* A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data* **1,** 140031 (2014).
4. Selevsek, N. *et al.* Reproducible and consistent quantification of the Saccharomyces cerevisiae proteome by SWATH-MS. *Mol. Cell. Proteomics* **14,** mcp.M113.035550–749 (2015).
5. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12,** 258–264 (2015).
6. 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).
7. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392 (2002).
8. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75,** 4646–4658 (2003).
9. 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).
10. 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* **4,** 923–925 (2007).
11. Reiter, L. *et al.* mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods* **8,** 430–435 (2011).
12. Shteynberg, D. *et al.* iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **10,** M111.007690 (2011).
13. Röst, H. L. *et al.* TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat. Methods* **13,** 777–783 (2016).
14. 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).
15. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805 (2011).
16. Zhang, Y. *et al.* ProteinInferencer: Confident protein identification and multiple experiment comparison for large scale proteomics projects. *J. Proteomics* **129,** 25–32 (2015).
17. Ochoa, A., Storey, J. D., Llinás, M. & Singh, M. Beyond the E-Value: Stratified Statistics for Protein Domain Prediction. *PLoS Comput. Biol.* **11,** e1004509 (2015).
18. Choi, H. & Nesvizhskii, A. I. Semisupervised Model-Based Validation of Peptide Identifications in Mass Spectrometry-Based Proteomics. *J. Proteome Res.* **7,** 254–265

(2008).
19. Shanmugam, A. K. & Nesvizhskii, A. I. Effective Leveraging of Targeted Search Spaces for Improving Peptide Identification in Tandem Mass Spectrometry Based Proteomics. *J. Proteome Res.* **14,** 5169–5178 (2015).
20. Wang, X. *et al.* Protein identification using customized protein sequence databases derived from RNA-Seq data. *J. Proteome Res.* **11,** 1009–1017 (2012).
21. Ning, K., Fermin, D. & Nesvizhskii, A. I. Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. *Proteomics* **10,** 2712–2718 (2010).
22. Tharakan, R., Edwards, N. & Graham, D. R. M. Data maximization by multipass analysis of protein mass spectra. *Proteomics* **10,** 1160–1171 (2010).
23. 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).
24. Noble, W. S. Mass spectrometrists should search only for peptides they care about. *Nat. Methods* **12,** 605–608 (2015).
25. Wang, J. *et al.* MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. *Nat. Methods* **12,** 1106–1108 (2015).
26. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100,** 9440–9445 (2003).
27. Collins, B. C. *et al. Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry*. *Nature communications* (2017). doi:(in press)
28. Liu, Y. *et al.* Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* **11,** 786–786 (2015).
29. 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).
30. Escher, C. *et al.* Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **12,** 1111–1121 (2012).
31. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J. R. Statist. Soc. B* **57,** 289–300 (1995).
32. 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 (2007).