How do Protein Contaminants Influence DDA and DIA Proteomics

Ashley M. Frankenfield[1], Jiawei Ni[1], Mustafa Ahmed[1], Ling Hao[1*]

[1]Department of Chemistry, The George Washington University, Science and Engineering Hall 4000, 800 22nd St., NW, Washington, DC, 20052, USA

*Corresponding author
Ling Hao
Assistant Professor
Department of Chemistry
George Washington University
Tel.: +1 (202) 994-4492
E-mail: linghao@gwu.edu

ABSTRACT

Mass spectrometry-based proteomics is constantly challenged by the presence of contaminant background signals. In particular, protein contaminants from reagents and sample handling are often abundant and impossible to avoid. For data-dependent acquisition (DDA) proteomics, exclusion list can be used to reduce the influence of protein contaminants. However, protein contamination has not been evaluated and is rarely addressed in data-independent acquisition (DIA). How protein contaminants influence proteomics data is also unclear. In this study, we established the protein contaminant FASTA and spectral libraries that are applicable to all proteomic workflows and evaluated the impact of protein contaminants on both DDA and DIA proteomics. We demonstrated that including our contaminant libraries can reduce false discoveries and increase protein identifications, without influencing the quantification accuracy in various proteomic software platforms. With the pressing need to standardize proteomic workflow in the research community, we highly recommend including our contaminant FASTA and spectral libraries in all bottom-up proteomics workflow. Our contaminant libraries and a step-by-step tutorial to incorporate these libraries in different DDA and DIA data analysis platforms can be a valuable resource for proteomics researchers, which are freely accessible at https://github.com/HaoGroup-ProtContLib.

KEYWORDS: protein contaminant, contamination, spectral library, FASTA, DIA, DDA, DIA-NN, Spectronaut, keratin, trypsin

INTRODUCTION

Mass spectrometry (MS)-based proteomics is constantly challenged by the exogenous contaminants and interferences that can be introduced into samples throughout the experimental workflow. Contaminations from polymers, detergents, solvents, ion sources, and other additives are often singly charged, which can be avoided by the careful selection of reagents or removed by ion mobility MS interface (*e.g.* FAIMS).[1–3] However, contaminant proteins and peptides are almost impossible to eliminate from proteomics workflows. For example, keratins from researchers' skin and hair can be found on all surfaces and dust during sample handling.[4] Rodent and sheep keratins can originate from animal facilities and wool clothing. Residue cell culture medium can lead to bovine protein contaminations. Protein digestion enzymes (*e.g.* trypsin and Lys-C) and the production of enzymes (from bacteria or bovine) can introduce protein contaminants into bottom-up proteomics workflow.[1] Additionally, bovine serum albumin (BSA), immobilized antibodies, and affinity tags (*e.g.* streptavidin, FLAG, HA) from affinity columns/beads also represent major contaminants in immunoassays and affinity purification MS.[5,6] These exogenous contaminant proteins/peptides can compete with real samples in the MS ion source, occupy the cycle times in the mass analyzer, reduce the number of useful peptide spectra, and hinder the detection of low abundant proteins from complex biological samples.

Due to the negative effects of protein contaminants in MS-proteomics, various methods have been implemented to combat this problem. Keratin contaminations can be reduced by using a laminar flow hood and fastidiously wiping down surfaces with ethanol and water.[7] However, it is almost impossible to eliminate keratins from proteomic experiments. Contaminations from digestion enzymes and affinity tags can be reduced by carefully optimizing the amount of enzymes and beads. But such practices may not be feasible for MS facilities and biological samples with limited amounts. Sample type-specific interferences have been evaluated previously, such as non-specific interactions in affinity purification and contaminations from plasma proteomics.[5,8] Note that these interference proteins are sample-specific and may actually be useful proteins in other proteomic experiments. Therefore, these interference proteins cannot be marked as exogenous contaminant proteins for all proteomic experiments. For data-dependent acquisition (DDA) proteomics, an exclusion list can be used to disregard specific ions from being isolated for MS/MS fragmentation.[9] However, exclusion list is specific to LC-MS gradient

and instrument, which is difficult to transfer across different MS platforms and laboratories. Peptides with similar *m/z* and retention times could also be accidentally excluded in complex biological samples.[9–12] Additionally, contaminant FASTA libraries can be used in DDA data analysis to mark and remove contaminant proteins from the dataset.[13] Although widely used, current contaminant FASTA files used in Mascot, MaxQuant[14], or other proteomics platforms have not been updated in years.

Despite various strategies to reduce the influence of protein contaminants in DDA proteomics, protein contamination has not been evaluated and is rarely addressed in data-independent acquisition (DIA) proteomics. Many exogenous contaminants from different species cannot be identified unless included in FASTA or spectral libraries. DDA exclusion list is not compatible with DIA because all co-eluting peptides within a pre-determined isolation window are fragmented together regardless of precursor intensities. Due to the wide isolation window in DIA, we hypothesized that contaminant proteins/peptides can be especially problematic if left unaddressed, leading to false identifications in DIA proteomics.

DIA data analysis can be conducted using spectral library-based software tools (*e.g.* OpenSWATH[15], Spectronaut[16], DIA-NN[17], Skyline[18], EncylopeDIA[19], MaxDIA[20]) or library-free strategies with in-silico digested pseudo peptide spectra based on FASTA protein sequences (*e.g.* DirectDIA[21], DIA-NN[17], DIA-Umpire[22], PECAN[23,24], DeepDIA[25]). While contaminant FASTA libraries are widely implemented in DDA data analysis, they have not been updated for years and are rarely used for DIA data analysis.[26] To our knowledge, there is no contaminant spectral library that is publicly available and universally applicable for DIA proteomics.

In this study, we created a series of contamination-only samples to establish contaminant protein spectral libraries. We generated a new contaminant FASTA library containing 381 protein contaminants commonly found in bottom-up proteomic experiments. We then evaluated how protein contaminants influence identification and quantification in DDA and DIA proteomics. The benefits and applicability of these contaminant libraries were demonstrated in various DDA and DIA data analysis platforms. These contaminant FASTA and spectral libraries are freely accessible at https://github.com/HaoGroup-ProtContLib with straightforward user manual to promote standardized and reproducible proteomics data analysis and reporting pipeline in the broad proteomics community.[27–29]

MATERIALS AND METHODS

Generation of Contaminant-Only Samples

We generated a series of contaminant-only samples by adding different proteolytic enzymes to the lysis buffer (1M Urea in 50 mM Tris-HCl) as well as commonly used beads coated with affinity tags. The proteolytic enzymes used here include sequencing grade Trypsin, Trypsin Gold, Trypsin/Lys-C, and Lys-C from Promega. The beads used here include Sero-Mag streptavidin magnetic beads (Cytivia), Anti-Flag M2 affinity agarose beads (Sigma), and EZview Red anti-HA affinity agarose beads (Sigma). Clean ungloved hands were purposely rubbed together above these samples to increase keratin contaminations.

Human Cell Culture and Mouse Brain Tissues

HEK293 cells were maintained in DMEM/F12 HEPES medium (Gibco) containing 10% of fetal bovine serum (FBS). Mouse brain samples were obtained from wild-type mice (C57/B6) under protocols approved by the George Washington University Institutional Animal Care and Use Committee. HEK cells and mouse brain samples were lysed in 8 M Urea, 50 mM Tris-HCl buffer and sonicated for 15 min in an ice-cold water bath using s QSonica Q700 Sonicator with alternating cycles of 1 min on and 30 s off. Protein lysates were clarified by 15 min of centrifugation at 12,000 rpm at 4 ºC and stored in -80 ºC. Total protein concentrations were determined using a detergent-compatible colorimetric protein assay (DCA, BioRad).

Proteomic Sample Preparation

The routine bottom-up proteomic workflow was conducted for contaminant-only samples, HEK cells and mouse brain lysates as described previously.[30,31] Briefly, disulfide bonds were reduced using 5 mM Tris(2-carboxylethyl)phosphine (TCEP) for 30 min, 15 mM of iodoacetamide for 30 min in dark, and 5 mM TCEP for 10 min on a ThermoMixer shaking at 1,200 rpm at 37 ºC. Protein digestions were conducted using various enzymes (contaminant-only samples) and Trypsin/Lys-C (HEK and mouse samples) for 18 hours at 37 ºC on a ThermoMixer, and quenched with 10% trifluoracetic acid until pH < 3. Peptides were then desalted on a Waters Oasis HLB Plate using the manufacturer's protocol, dried down under SpeedVac, and stored at -30 ºC.

LC-MS/MS Analysis for DDA and DIA Proteomics

Peptide samples were analyzed on a Dionex UltiMate 3000 RSLCnano system coupled with a Thermo Fisher Q-Exactive HF-X mass spectrometer. The mobile phase buffer A was 0.1% formic acid in water, and buffer B was 0.1% formic acid in acetonitrile. HEK cells and mouse brain samples were injected onto an Acclaim PepMAP C18 trap column (3 µm, 100Å, 75 µm × 2cm) and further separated on an Easy-spray PepMap C18 column (2 µm, 100Å, 75 µm × 75cm) with a flow rate of 0.2 µL/min, an LC gradient of 210 min, and a column temperature of 55 ºC. Contaminant-only samples were analyzed with a 15 cm PepMap C18 column, a flow rate of 0.3 µL/min, and an LC gradient of 120 min. For DDA analysis, MS scans from *m/z* 380 to 1,500 with a resolving power of 120K (at *m/z* 200 FWHM), an automatic gain control (AGC) target of 1×10^6, and a maximum injection time (maxIT) of 50 ms. Precursors were isolated at a window of *m/z* 1.4 and fragmented with a normalized collision energy (NCE) of 30%, a resolving power of 7.5K for MS/MS, and a maxIT of 40 ms. For DIA analysis, MS scans from *m/z* 400 to 1000 at a resolving power of 60K, an AGC target of 1×10^6, and a maxIT of 30 ms. The precursor isolation window was set to *m/z* 8.0 (staggered) with 75 sequential DIA MS/MS scans between *m/z* 400 to 1000 at a resolving power of 30K, an AGC target of 5×10^5, a MaxIT of 30 ms, and an NCE of 30%.

Repository Data from ProteomeXchange

Two repository datasets from the ProteomeXchange website were downloaded and reanalyzed using our contaminant libraries. Repository dataset A is a HepG2 human cell DIA dataset (PXD022589) containing 27 raw data.[20] Dataset B is a fractionated mouse cortex DIA dataset (PXD005573) containing 12 raw data.[32] Additionally, a fractionated HEK and HeLa cell DDA dataset (PXD001468) was used to generate a spectral library for library-based DIA data analysis.[33]

DDA Proteomics Data Analysis

Common contaminant protein sequence library (*e.g.* enzymes, keratins, affinity tags, bovine proteins) were created by updating existing contaminant lists online, searching new contaminants on Uniprot website, and combining them into a new FASTA library (**Supporting Information FASTA and Table S1**). All DDA proteomic datasets in this study were analyzed with both the

MaxQuant (2.0.2.0) and Thermo Fisher Proteome Discoverer (2.4.1.15) software with similar parameters. Contaminant-only samples were analyzed with the contaminant FASTA library only. HEK cells and mouse brain samples were analyzed using the Swiss-Prot Homo sapiens database (reviewed) and mus musculus database (reviewed), respectively, with and without our contaminant FASTA library. The false discovery rate for protein and peptide spectral matches (PSMs) identifications was set at a false discovery rate (FDR) of 0.01. Trypsin or LysC enzyme was used with a maximum missed cleavage of two. Precursor tolerance was set to 20 ppm. The fixed modification was cysteine carbamidomethyl, and variable modifications were methionine oxidation and protein N-terminus acetylation.

DIA Proteomics Data Analysis

Spectronaut software: A contaminant-only spectral library was generated using the set of contaminant-only DDA proteomic dataset using the Pulsar search engine in Spectronaut 15.[34] Mouse and human spectral libraries were also generated using Pulsar. Two spectral libraries were generated for each sample type with and without including the contaminant-only samples in Pulsar (**Supplemental Table S2**). The "Library Generation Step" of Pulsar was conducted using "BGS Factory Settings". Specific trypsin digestion was set with a maximum of two missed cleavages. A fixed carbamidomethyl modification of cysteine, and up to three variable modifications for oxidation of methionine and acetylation of the protein N-terminus were allowed. PSM, peptide and protein FDR were set to 0.01. Both library-based and library-free (DirectDIA) analyses were performed in Spectronaut 15 using default settings. The quantification step was modified to perform an interference correction that used only identified peptides to train the machine-learning model. No cross-run normalization or imputation was used.

DIA-NN software: Both spectral library-based and library-free DIA analyses were performed in DIA-NN (v1.8).[17] Raw data files were converted to the open-format *.mzML* using the msConvert feature of the ProteoWizard package.[35] Library-based analysis was conducted in DIA-NN using the spectral libraries established above in Spectronaut Pulsar. All settings were default and the same as in the Spectronaut. For library-free analysis, FASTA digest was selected. The spectral libraries were also included to train the deep learning model.

Post-Data Analysis Filtering

To increase the confidence of protein/peptide identifications, proteins that were identified with only one precursor or an intensity below 10 were removed from all datasets using R. Contaminant proteins can be easily filtered out from the results by searching the "Cont_"prefix in the UniProt protein group identifier from the results. Contaminant proteins were removed before calculating the coefficient of variation and Spearman's correlation for the proteomic datasets.

Data Availability

All raw files have been deposited to the ProteomeXchange Consortium with the data identifier, PXD031139. The protein contaminant library and step-by-step user tutorial is also freely accessible at https://github.com/HaoGroup-ProtContLib.

RESULTS AND DISCUSSION

Building the Contaminant Protein FASTA and Spectral Libraries

Exogenous contaminant proteins orignated from reagents and sample handling are mostly shared in all bottom-up proteomic experiments. Therefore, we aim to build exogenous contaminant protein libraries that can be used in all bottom-up proteomics (**Figure 1)**. Although widely used for DDA proteomics, the list of common protein contaminants from Mascot and Maxquant platforms have not been updated for years and contained many incorrect Uniprot IDs. Some sample-specific interference proteins were also incorrectly listed as contaminant proteins. Therefore, we first built a new contaminant FASTA library by manually merging several contaminant lists online, updating their Uniprot entry IDs, deleting noncontaminant proteins, searching new contaminant proteins on Uniprot, and combining them into a new FASTA file. Our new contaminant FASTA library contains 381 contaminant proteins including all human keratins and skin-derived proteins, common bovine contaminants from cell culture and affinity columns, various proteolytic enzymes, affinity tags, and other contaminants (**Supplemental FASTA and Table S1**). When compared to the MaxQuant contaminant list, our new FASTA

library is up-to-date for all Uniprot IDs and contains an additional 183 contaminant proteins (**Figure 2A**). This new FASTA library can be used for both DDA and DIA proteomics. We also added a "Cont_" prefix in each contaminant entry in the FASTA library, allowing contaminant proteins to be easily filtered and removed in the result files.

To establish comprehensive contaminant protein spectral libraries for DIA proteomics, we created a series of contaminant-only samples using various proteolytic enzymes and affinity purification beads. We validated the presence of each contaminant peptides by creating spectral libraries in MaxQuant, Proteome Discoverer and Spectronaut Pulsar (**Figure 1**). Hundreds of contaminant peptides were detected throughout the LC-MS gradient (**Figure 2B and Supplemental Table S3)**. Since trypsin and LysC are the two most used enzymes for bottom-up proteomics, we created two DIA spectral libraries using Spectronaut Pulsar: one for tryptic contaminant peptides, one for LysC-digested contaminant peptides. These spectral libraries are built from highly confident fragment ions assigned to each peptide sequence (**Figure 2C, Supplemental Table S4**), also freely accessible on ProteomeXchange (PXD031139). We further examined the contaminant datasets and found that human keratins and Lys-C enzyme produce the largest number of contaminant PSMs **(Supplemental Table S3)**. LysC enzyme provides higher cleavage efficiency at lysine and is therefore often used in combination with trypsin to improve digestion efficiency.[36] However, LysC enzyme contains almost two fold more arginine/lysine residues compared to trypsin and can therefore generate numerous contaminant peptides as shown in **Figure 2**. As expected, the quality of enzymes also influences the number of contaminant peptides. Trypsin Gold produced less contaminant PSMs compared to sequencing grade trypsin. Additionally, bovine protein contaminants (albumin *et. al.*) were identified in all affinity purification beads despite conducting pre-washing steps. Streptavidin coated beads generated overwhelming streptavidin peptide signals which is consistent with our previous findings.[6] These exogenous contaminant proteins are often originated from a different species which will not be identified unless the contaminant FASTA library is included in the data analysis workflow.

Contaminant Peptides can Cause False Discoveries in DIA Proteomics

Contaminant FASTA library has been widely used for DDA proteomics, but is rarely included in DIA data analysis.[26,30,37,38] Since DIA uses a much wider precursor isolation window (4-15 Da) compared to DDA (0.4-2 Da), contaminant peptides in DIA are more likely to be coeluted and co-fragmented with other peptides. If not addressed properly, contaminant peptides can cause false identifications of peptides/proteins. To evaluate the influence of contaminant peptides, we analyzed several DIA proteomic datasets with and without our contaminant FASTA library. As shown in **Figure 3A**, when the contaminant FASTA library is not included for data analysis, a contaminant Lys-C peptide was misidentified as a KIF20B peptide due to numerous shared peptide fragments. After including the contaminant library, the peak picking algorithm identified an additional y3 ion and assigned the fragmentation spectra to Lys-C instead of KIF20B with a higher confidence and lower peptide *q*-values. Similar scenario happened to another bovine contaminant protein SERPINA1 which was misidentified as CFAP100 (**Figure 3B**). Including the contaminant library allows the identification of three additional fragments and the correct assignment to SERPINA1 contaminant peptide. Furthermore, as contaminant peptides elute throughout the LC gradient and mass range (**Figure 2B**), many contaminant peptides can be coeluted and co-fragmented with real peptides of interest. **Figure 3C** and **3D** showed examples of coeluted and co-fragmented contaminant peptides in our DIA datasets with an 8 Da isolation window. When larger isolation windows are used such as 10 Da or 12 Da, more peptides including contaminants will be co-fragmented to generate highly convoluted MS/MS spectra.[39,40] Contaminant peptides with high abundance can also suppress the detection of low abundant peptides by competing with them in the ion source and mass analyzer. Therefore, including a contaminant library can greatly reduce false discoveries in DIA data, particularly for complex biological samples. However, careful optimization of experimental workflow and DIA parameters are still crucial and fundamental to reduce the numbers and abundances of contaminant signals.

Including Contaminant Protein Library Improves both DDA and DIA Proteomics

Contaminant libraries can be integrated into the DDA and DIA data analysis workflow via different strategies. DDA and library-free DIA analyses only requires the contaminant FASTA protein sequences. But library-based DIA analysis requires both FASTA and spectral libraries.

Contaminant spectral library can be generated in two ways: 1) an integrated spectral library built from contaminant-only raw data and custom proteomics data together; 2) two separate spectral libraries for contaminant and custom proteomics data. Contaminant FASTA file is also required when building these spectral libraries. In Spectronaut software, multiple spectral libraries can be included during data analysis. We found that the integrated spectral library performs similar with two separate libraries with slightly higher total protein identifications in some datasets (**Supplemental Figure S1**). Either method is better compared to the results analyzed without the contaminant library. However, many other DIA software platforms do not allow the inclusion of multiple spectral libraries, and thus require an integrated spectral library. Fortunately, including the additional contaminant FASTA and spectral libraries did not increase the software processing time for multiple DDA (Proteome Discoverer, Maxquant) and DIA (DIA-NN, Spectronaut) platforms.

To demonstrate the benefits of contaminant protein libraries for both DDA and DIA proteomics, HEK cells and mouse brain samples were analyzed in DDA and DIA workflows in various data analysis software (**Figure 4**). The overall increase of protein and peptide identifications were around 0.9% and 1.3%, respectively across all software and sample types. After removing the contaminants, more peptide IDs and similar protein IDs were achieved when analyzed with the contaminant library. For various DIA platforms, library-free DIA-NN generated the highest number of protein and peptide IDs possibly due to the deep learning model implemented in search algorithm. Besides in-house generated proteomics data, we also analyzed repository datasets with and without contaminant library. Two DIA repository datasets were downloaded from ProteomeXchange: repository dataset A from HepG2 human cell samples[20] and dataset B from mouse brain samples[32]. Higher number of proteins and peptides were identified when the data was analyzed with contaminant libraries (**Figure 5**). Particularly, after removing contaminants, almost 5% more proteins and peptides were identified when the data was analyzed with contaminant libraries in Spectronaut library-based platform. Benefited from the additional contaminant spectral library, library-based DIA platforms provided greater increase of identifications compared to library-free platforms. Many bovine contaminant proteins were identified from repository dataset A similar with our in-house generated HEK cell dataset, which can be traced back to the fetal bovine serum (FBS) used for human cell culture. To

minimize the contaminations from cell culture medium, we highly recommend two to three times of PBS washes during cell harvest.

Since our contaminant libraries can improve protein/peptide identifications, we further assessed protein quantification with and without contaminant libraries. Coefficient of variation (CV) of all quantified proteins from HEK cells (**Figure 6A**) and mouse brain samples (**Figure 6B**) were calculated after removing the contaminant proteins. No significant differences were observed with and without including contaminant libraries. DIA-NN resulted more protein identifications but higher CVs compared to Spectronaut platform. Library-based methods provided less variation and better reproducibility compared to library-free methods, consistent with other reported studies.[19,41] Protein intensities were not exactly the same when the data was analyzed with or without contaminant libraries, but they did correlate very well with spearman's correlation close to 1 (**Figure 6C and 6D**). These results demonstrated that contaminant libraries didn't influence DIA protein quantification.

CONCLUSIONS

This study filled a critical gap in bottom-up proteomics by establishing and evaluating contaminant protein libraries to reduce false discoveries and improve identifications in both DDA and DIA proteomics. Although the software used here (Spectronaut, DIA-NN, Maxquant, Proteome Discoverer) are not an exhaustive list of all available data analysis platforms, we believe that our contaminant libraries can be universally applied to all bottom-up DIA and DDA proteomics software. In fact, we provided step-by-step tutorial on how to best incorporate our contaminant FASTA and Spectral libraries for many other software platforms such as Skyline[18], MaxDIA[20], and PECAN[23] (**Supplemental Tutorial**). We will also continue updating and enriching our contaminant libraries to include sample type-specific contaminant libraries. These freely accesible contaminant FASTA and spectral libraries can be valuable resources for proteomic researchers and facilitate the standardization of proteomic data analysis across different laboratories.

SUPPLEMENTAL INFORMATION

Supplemental FASTA. Contaminant protein FASTA with Cont_ prefix.

Supplemental Tutorial. Tutorial for using contaminant libraries for DDA and DIA data analysis in various proteomics software.

Supplemental Figure S1. Evaluation of different methods to build contaminant spectral libraries in Spectronaut software.

Supplemental Table S1: Contaminant protein information in the FASTA library and potential source of contaminations.

Supplemental Table S2: List of the established FASTA and Pulsar spectral libraries.

Supplemental Table S3: Summary of the protein contaminants identified in contaminant-only samples.

Supplemental Table S4: List of the contaminant peptides and fragments in Pulsar spectral libraries.

AUTHOR CONTRIBUTIONS

A.M.F. and L.H. designed the study and wrote the manuscript with inputs and revisions from all coauthors. A.M.F, J.N., and A.M. conducted the experiments. A.M.F performed data analysis. All authors have read and agreed to the published version of this manuscript.

ACKNOWLEDGEMENTS

This study is supported by the NIH grant (R01NS121608). L.H acknowledges the ORAU Ralph E. Powe Junior Faculty Enhancement Award. A.M.F acknowledges the ARCS-Metro Washington Chapter Scholarship and the Bourbon F. Scribner Endowment Fellowship. We thank the Vertes lab and Lu Lab at GW for the access to the SpeedVac instrument and mouse brain samples.

CONFLICTS OF INTEREST

The authors declare no competing financial interests.

REFERENCES

(1) Keller, B. O.; Sui, J.; Young, A. B.; Whittal, R. M. Interferences and Contaminants Encountered in Modern Mass Spectrometry. *Anal. Chim. Acta* **2008**, *627* (1), 71–81. https://doi.org/10.1016/j.aca.2008.04.043.

(2) Rardin, M. J. Rapid Assessment of Contaminants and Interferences in Mass Spectrometry Data Using Skyline. *J. Am. Soc. Mass Spectrom.* **2018**, *29* (6), 1327–1330. https://doi.org/10.1007/s13361-018-1940-z.

(3) Guevremont, R. High-Field Asymmetric Waveform Ion Mobility Spectrometry: A New Tool for Mass Spectrometry. *J. Chromatogr. A* **2004**, *1058* (1–2), 3–19.

(4) Fox, K.; Castanha, E.; Fox, A.; Feigley, C.; Salzberg, D. Human K10 Epithelial Keratin Is the Most Abundant Protein in Airborne Dust of Both Occupied and Unoccupied School Rooms. *J. Environ. Monit.* **2008**, *10* (1), 55–59.

(5) Mellacheruvu, D.; Wright, Z.; Couzens, A. L.; Lambert, J. P.; St-Denis, N. A.; Li, T.; Miteva, Y. V.; Hauri, S.; Sardiu, M. E.; Low, T. Y.; Halim, V. A.; Bagshaw, R. D.; Hubner, N. C.; Al-Hakim, A.; Bouchard, A.; Faubert, D.; Fermin, D.; Dunham, W. H.; Goudreault, M.; Lin, Z. Y.; Badillo, B. G.; Pawson, T.; Durocher, D.; Coulombe, B.; Aebersold, R.; Superti-Furga, G.; Colinge, J.; Heck, A. J. R.; Choi, H.; Gstaiger, M.; Mohammed, S.; Cristea, I. M.; Bennett, K. L.; Washburn, M. P.; Raught, B.; Ewing, R. M.; Gingras, A. C.; Nesvizhskii, A. I. The CRAPome: A Contaminant Repository for Affinity Purification-Mass Spectrometry Data. *Nat. Methods* **2013**, *10* (8), 730–736. https://doi.org/10.1038/nmeth.2557.

(6) Frankenfield, A. M.; Fernandopulle, M. S.; Hasan, S.; Ward, M. E.; Hao, L. Development and Comparative Evaluation of Endolysosomal Proximity Labeling-Based Proteomic Methods in Human IPSC-Derived Neurons. *Anal. Chem.* **2020**, *92* (23), 15437–15444. https://doi.org/10.1021/acs.analchem.0c03107.

(7) Biringer, R.; San Parkway, R. O. Protocol for a Keratin-Free Environment. *Thermo Electron Corp. San Jose, CA* **2002**.

(8) Geyer, P. E.; Voytik, E.; Treit, P. V; Doll, S.; Kleinhempel, A.; Niu, L.; Müller, J. B.; Buchholtz, M.; Bader, J. M.; Teupser, D.; Holdt, L. M.; Mann, M. Plasma Proteome Profiling to Detect and Avoid Sample□related Biases in Biomarker Studies. *EMBO Mol. Med.* **2019**, *11* (11), 1–12. https://doi.org/10.15252/emmm.201910427.

(9) Hodge, K.; Have, S. Ten; Hutton, L.; Lamond, A. I. Cleaning up the Masses: Exclusion Lists to Reduce Contamination with HPLC-MS/MS. *J. Proteomics* **2013**, *88*, 92–103. https://doi.org/10.1016/j.jprot.2013.02.023.

(10) Schenk, S.; Schoenhals, G. J.; de Souza, G.; Mann, M. A High Confidence, Manually Validated Human Blood Plasma Protein Reference Set. *BMC Med. Genomics* **2008**, *1* (1), 41. https://doi.org/10.1186/1755-8794-1-41.

(11) Hebert, A. S.; Thöing, C.; Riley, N. M.; Kwiecien, N. W.; Shiskova, E.; Huguet, R.; Cardasis, H. L.; Kuehn, A.; Eliuk, S.; Zabrouskov, V.; Westphall, M. S.; McAlister, G. C.; Coon, J. J. Improved Precursor Characterization for Data-Dependent Mass Spectrometry. *Anal. Chem.* **2018**, *90* (3), 2333–2340. https://doi.org/10.1021/acs.analchem.7b04808.

(12) Hiemstra, T. F.; Charles, P. D.; Hester, S. S.; Karet, F. E.; Lilley, K. S. Uromodulin Exclusion List Improves Urinary Exosomal Protein Identification. *J. Biomol. Tech.* **2011**, *22* (4), 136–145.

(13) Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M. Quantifying the Impact of Chimera MS/MS Spectra on Peptide Identification in Large-Scale Proteomics Studies. *J. Proteome Res.* **2010**, *9* (8), 4152–4160. https://doi.org/10.1021/pr1003856.

(14) Tyanova, S.; Temu, T.; Cox, J. The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics. *Nat Protoc* **2016**, *11* (12), 2301–2319. https://doi.org/10.1038/nprot.2016.136.

(15) Röst, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinoviä, S. M.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Malmström, J.; Malmström, L.; Aebersold, R. OpenSWATH Enables Automated, Targeted Analysis of Data-Independent Acquisition MS Data. *Nat. Biotechnol.* **2014**, *32* (3), 219–223. https://doi.org/10.1038/nbt.2841.

(16) Bruderer, R.; Bernhardt, O. M.; Gandhi, T.; Miladinović, S. M.; Cheng, L. Y.; Messner, S.; Ehrenberger, T.; Zanotelli, V.; Butscheid, Y.; Escher, C.; Vitek, O.; Rinner, O.; Reiter, L. Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. *Mol. Cell. Proteomics* **2015**, *14* (5), 1400–1410. https://doi.org/10.1074/mcp.M114.044305.

(17) Demichev, V.; Messner, C. B.; Vernardis, S. I.; Lilley, K. S.; Ralser, M. DIA-NN: Neural Networks and Interference Correction Enable Deep Proteome Coverage in High Throughput. *Nat. Methods* **2020**, *17* (1), 41–44. https://doi.org/10.1038/s41592-019-0638-x.

(18) Egertson, J. D.; MacLean, B.; Johnson, R.; Xuan, Y.; MacCoss, M. J. Multiplexed Peptide Analysis Using Data-Independent Acquisition and Skyline. *Nat. Protoc.* **2015**, *10* (6), 887–903. https://doi.org/10.1038/nprot.2015.055.

(19) Searle, B. C.; Pino, L. K.; Egertson, J. D.; Ting, Y. S.; Lawrence, R. T.; MacLean, B. X.; Villén, J.; MacCoss, M. J. Chromatogram Libraries Improve Peptide Detection and Quantification by Data Independent Acquisition Mass Spectrometry. *Nat. Commun.* **2018**, *9* (1). https://doi.org/10.1038/s41467-018-07454-w.

(20) Sinitcyn, P.; Hamzeiy, H.; Salinas Soto, F.; Itzhak, D.; McCarthy, F.; Wichmann, C.; Steger, M.; Ohmayer, U.; Distler, U.; Kaspar-Schoenefeld, S.; Prianichnikov, N.; Yılmaz, Ş.; Rudolph, J. D.; Tenzer, S.; Perez-Riverol, Y.; Nagaraj, N.; Humphrey, S. J.; Cox, J. MaxDIA Enables Library-Based and Library-Free Data-Independent Acquisition

Proteomics. *Nat. Biotechnol.* **2021**. https://doi.org/10.1038/s41587-021-00968-7.

(21) Muntel, J.; Kirkpatrick, J.; Bruderer, R.; Huang, T.; Vitek, O.; Ori, A.; Reiter, L. Comparison of Protein Quantification in a Complex Background by DIA and TMT Workflows with Fixed Instrument Time. *J. Proteome Res.* **2019**, *18* (3), 1340–1351. https://doi.org/10.1021/acs.jproteome.8b00898.

(22) Tsou, C. C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A. C.; Nesvizhskii, A. I. DIA-Umpire: Comprehensive Computational Framework for Data-Independent Acquisition Proteomics. *Nat. Methods* **2015**, *12* (3), 258–264. https://doi.org/10.1038/nmeth.3255.

(23) Ting, Y. S.; Egertson, J. D.; Bollinger, J. G.; Searle, B. C.; Payne, S. H.; Noble, W. S.; MacCoss, M. J. PECAN: Library-Free Peptide Detection for Data-Independent Acquisition Tandem Mass Spectrometry Data. *Nat. Methods* **2017**, *14* (9), 903–908. https://doi.org/10.1038/nmeth.4390.

(24) Pino, L. K.; Just, S. C.; MacCoss, M. J.; Searle, B. C. Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries. *Mol. Cell. Proteomics* **2020**, *19* (7), 1088–1103. https://doi.org/10.1074/mcp.P119.001913.

(25) Yang, Y.; Liu, X.; Shen, C.; Lin, Y.; Yang, P.; Qiao, L. In Silico Spectral Libraries by Deep Learning Facilitate Data-Independent Acquisition Proteomics. *Nat. Commun.* **2020**, *11* (1), 1–11. https://doi.org/10.1038/s41467-019-13866-z.

(26) Mosen, P.; Sanner, A.; Singh, J.; Winter, D. Targeted Quantification of the Lysosomal Proteome in Complex Samples. *Proteomes* **2021**, *9* (1), 1–17. https://doi.org/10.3390/proteomes9010004.

(27) Taylor, C. F.; Paton, N. W.; Lilley, K. S.; Binz, P.-A.; Julian, R. K.; Jones, A. R.; Zhu, W.; Apweiler, R.; Aebersold, R.; Deutsch, E. W.; Dunn, M. J.; Heck, A. J. R.; Leitner, A.; Macht, M.; Mann, M.; Martens, L.; Neubert, T. A.; Patterson, S. D.; Ping, P.; Seymour, S. L.; Souda, P.; Tsugita, A.; Vandekerckhove, J.; Vondriska, T. M.; Whitelegge, J. P.; Wilkins, M. R.; Xenarios, I.; Yates, J. R.; Hermjakob, H. The Minimum Information about a Proteomics Experiment (MIAPE). *Nat. Biotechnol.* **2007**, *25* (8), 887–893. https://doi.org/10.1038/nbt1329.

(28) Neely, B.; Kirkpatrick, J.; Midha, M.; Stemmer, P.; Martin, L.; Wang, Y.; Herring, L.; Phinney, B.; Shan, B.; Chien, A. Data Independent Acquisition in Practice: Foundations and Resources for Implementing DIA. *J. Biomol. Tech. JBT* **2019**, *30* (Suppl), S46.

(29) Poulos, R. C.; Hains, P. G.; Shah, R.; Lucas, N.; Xavier, D.; Manda, S. S.; Anees, A.; Koh, J.; Mahboob, S.; Wittman, M. Strategies to Enable Large-Scale Proteomics for Reproducible Research. *Nat. Commun.* **2020**, *11* (1), 1–13.

(30) Li, H.; Uittenbogaard, M.; Navarro, R.; Ahmed, M.; Gropman, A.; Chiaramello, A.; Hao, L. Integrated Proteomic and Metabolomic Analyses of the Mitochondrial Neurodegenerative Disease MELAS. *Mol. Omi.* **2022**.

https://doi.org/10.1039/d1mo00416f.

(31) Thomas, S.; Hao, L.; DeLaney, K.; McLean, D.; Steinke, L.; Marker, P. C.; Vezina, C. M.; Li, L.; Ricke, W. A. Spatiotemporal Proteomics Reveals the Molecular Consequences of Hormone Treatment in a Mouse Model of Lower Urinary Tract Dysfunction. *J. Proteome Res.* **2020**.

(32) Bruderer, R.; Bernhardt, O. M.; Gandhi, T.; Xuan, Y.; Sondermann, J.; Schmidt, M.; Gomez-Varela, D.; Reiter, L. Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Mol. Cell. Proteomics* **2017**, *16* (12), 2296–2309. https://doi.org/10.1074/mcp.RA117.000314.

(33) Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P.; Biotechnol, N. An Ultra-Tolerant Database Search Reveals That a Myriad of Modified Peptides Contributes to Unassigned Spectra in Shotgun Proteomics HHS Public Access Author Manuscript. *Nat Biotechnol* **2015**, *33* (7), 743–749. https://doi.org/10.1038/nbt.3267.An.

(34) Bruderer, R.; Bernhardt, O. M.; Gandhi, T.; Miladinović, S. M.; Cheng, L. Y.; Messner, S.; Ehrenberger, T.; Zanotelli, V.; Butscheid, Y.; Escher, C.; Vitek, O.; Rinner, O.; Reiter, L. Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. *Mol. Cell. Proteomics* **2015**. https://doi.org/10.1074/mcp.M114.044305.

(35) Chambers, M. C.; MacLean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nat. Biotechnol.* **2012**, *30* (10), 918–920. https://doi.org/10.1038/nbt.2377.

(36) Saveliev, S.; Bratz, M.; Zubarev, R.; Szapacs, M.; Budamgunta, H.; Urh, M. Trypsin/Lys-C Protease Mix for Enhanced Protein Mass Spectrometry Analysis. *Nat. Methods* **2013**, *10* (11), i–ii. https://doi.org/10.1038/nmeth.f.371.

(37) Ghesquière, B.; Helsens, K.; Vandekerckhove, J.; Gevaert, K. A Stringent Approach to Improve the Quality of Nitrotyrosine Peptide Identifications. *Proteomics* **2011**, *11* (6), 1094–1098. https://doi.org/10.1002/pmic.201000526.

(38) Wang, J.; Bournes, P. E.; Bandeiras, N. Peptide Identification by Database Search of Mixture Tandem Mass Spectra. *Mol. Cell. Proteomics* **2011**, *10* (12), 1–11. https://doi.org/10.1074/mcp.M111.010017.

(39) Doellinger, J.; Blumenscheit, C.; Schneider, A.; Lasch, P. Isolation Window Optimization

of Data-Independent Acquisition Using Predicted Libraries for Deep and Accurate Proteome Profiling. *Anal. Chem.* **2020**, *92* (18), 12185–12192. https://doi.org/10.1021/acs.analchem.0c00994.

(40) Li, W.; Chi, H.; Salovska, B.; Wu, C.; Sun, L.; Rosenberger, G.; Liu, Y. Assessing the Relationship Between Mass Window Width and Retention Time Scheduling on Protein Coverage for Data-Independent Acquisition. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (8), 1396–1405. https://doi.org/10.1007/s13361-019-02243-1.

(41) Fernández-Costa, C.; Martínez-Bartolomé, S.; McClatchy, D. B.; Saviola, A. J.; Yu, N. K.; Yates, J. R. Impact of the Identification Strategy on the Reproducibility of the DDA and DIA Results. *J. Proteome Res.* **2020**, *19* (8), 3153–3161. https://doi.org/10.1021/acs.jproteome.0c00153.

Figures and Figure Legends

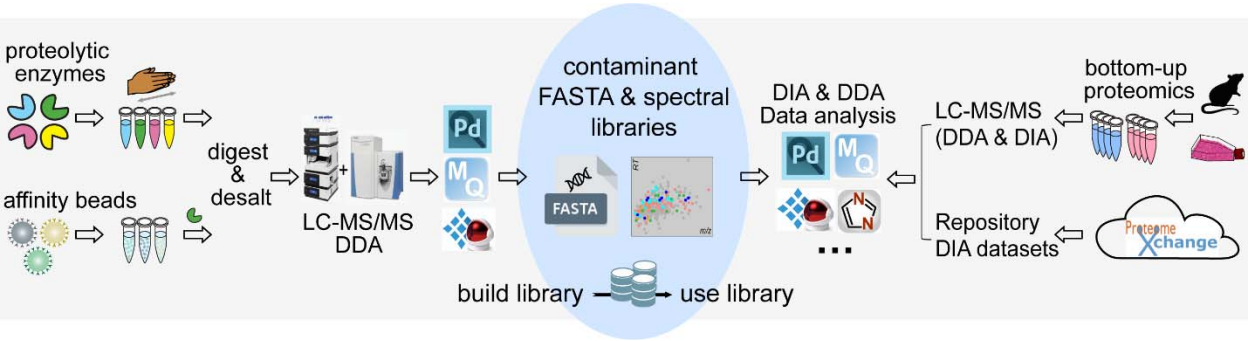

Figure 1: Schematic of Building and Using the Contaminant Libraries for DDA and DIA Proteomics. A series of contaminant-only samples were created by adding different proteolytic enzymes to keratin-contaminated lysis buffer and commonly used beads coated with affinity tags. DDA proteomic data of contaminant-only samples and updated contaminant FASTA protein sequences were combined to build the contaminant spectral libraries. Different biological samples and repository datasets were used to evaluate the contaminant FASTA and spectral libraries in various DDA and DIA software platforms.

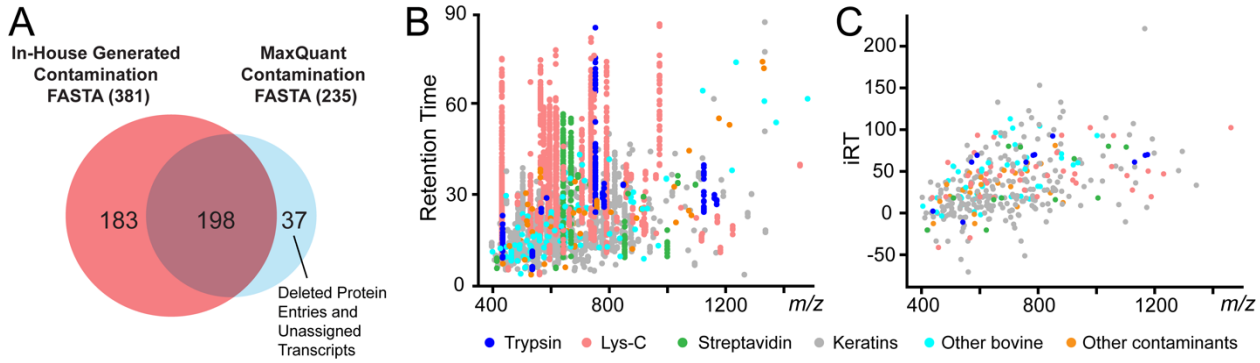

Figure 2: Characterization of the Contaminant Protein FASTA and Spectral Libraries. (A) Venn Diagram comparison of protein IDs from our newly generated contaminant FASTA library and the MaxQuant contaminant FASTA library. (B) Scatterplot showing contaminant peptides identified and merged from our custom contaminant-only samples in DDA LC-MS/MS. (C) Scatterplot showing contaminant peptides in our contaminant spectral libraries, generated by the Spectronaut Pulsar. iRT stands for in-silico normalized retention time.

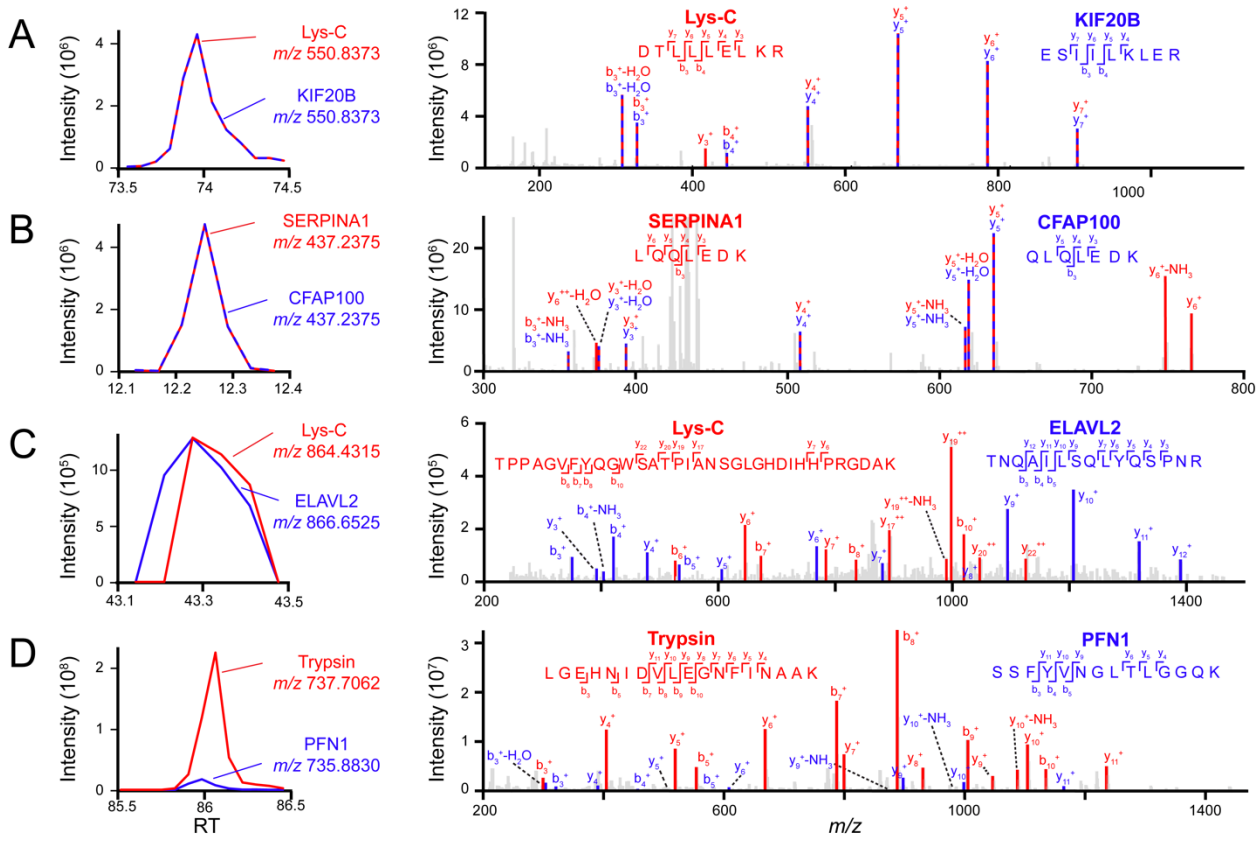

Figure 3: Examples of Misidentified and Coeluted Contaminant Peptides When Contaminant Libraries were not Included for Data Analysis. Contaminant peptide chromatograms and MS/MS fragments were shown in red, real peptides of interest were shown in blue. (A) Lys-C (enzyme contaminant) was misidentified as KIF20B. (B) SERPINA1 (bovine contaminant) was misidentified as CFAP100. (C) Lys-C was coeluted and co-fragmented with ELAVL2. (D) Trypsin was coeluted and co-fragmented with PFN1.

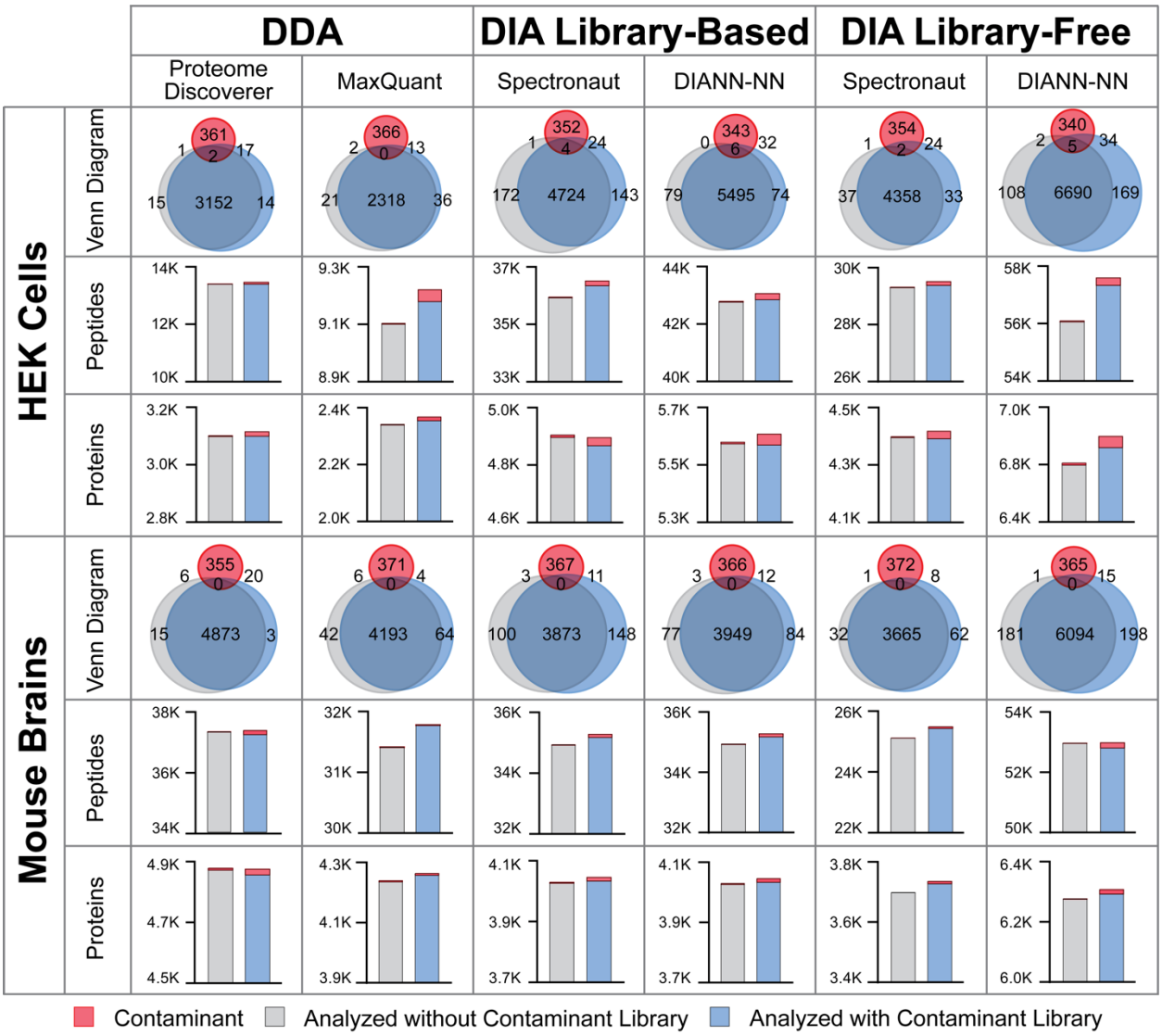

Figure 4: **Evaluation of Protein/Peptide Identifications Influenced by the Protein Contaminant Libraries in DDA and DIA Proteomics.** HEK cells and mouse brain samples were analyzed by various DDA and DIA software platforms, with (blue) and without contaminant libraries (grey). Contaminant proteins were marked in red.

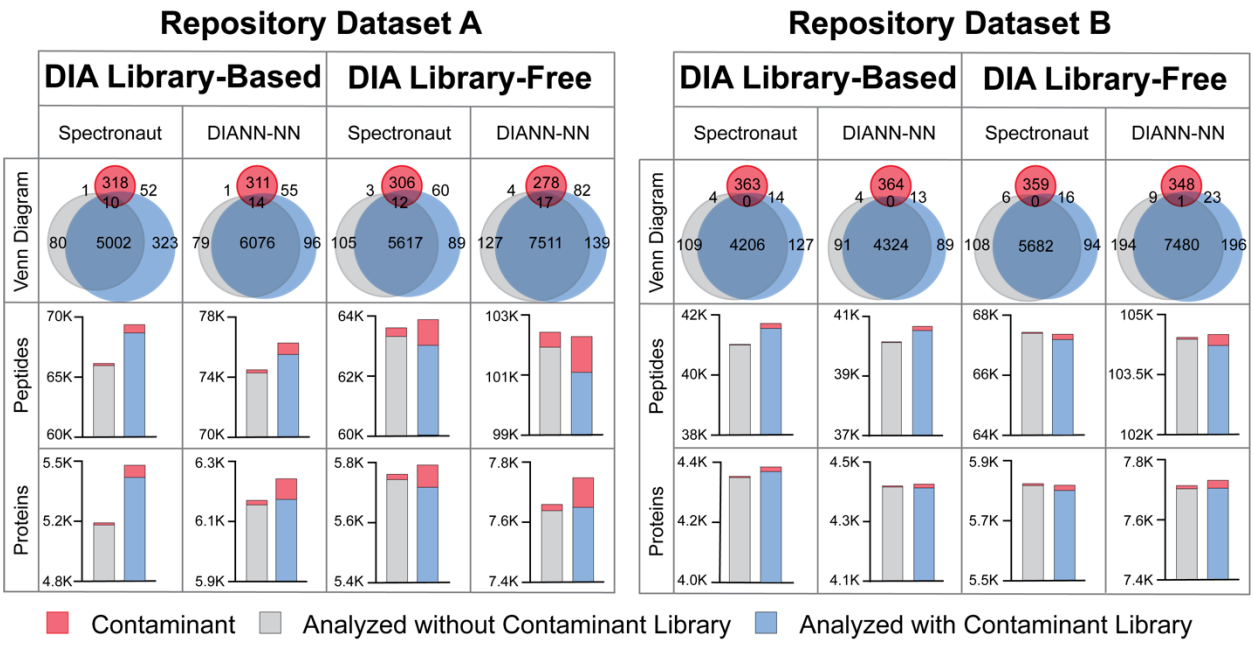

Figure 5: Repository DIA Proteomic Datasets Reanalyzed with and without the Contaminant Protein Libraries. Various DIA data analysis platforms were used to reanalyze two DIA datasets (Human cells: PXD022589; mouse cortex: PXD005573) with (blue) and without contaminant libraries (grey). Contaminant proteins were marked in red.

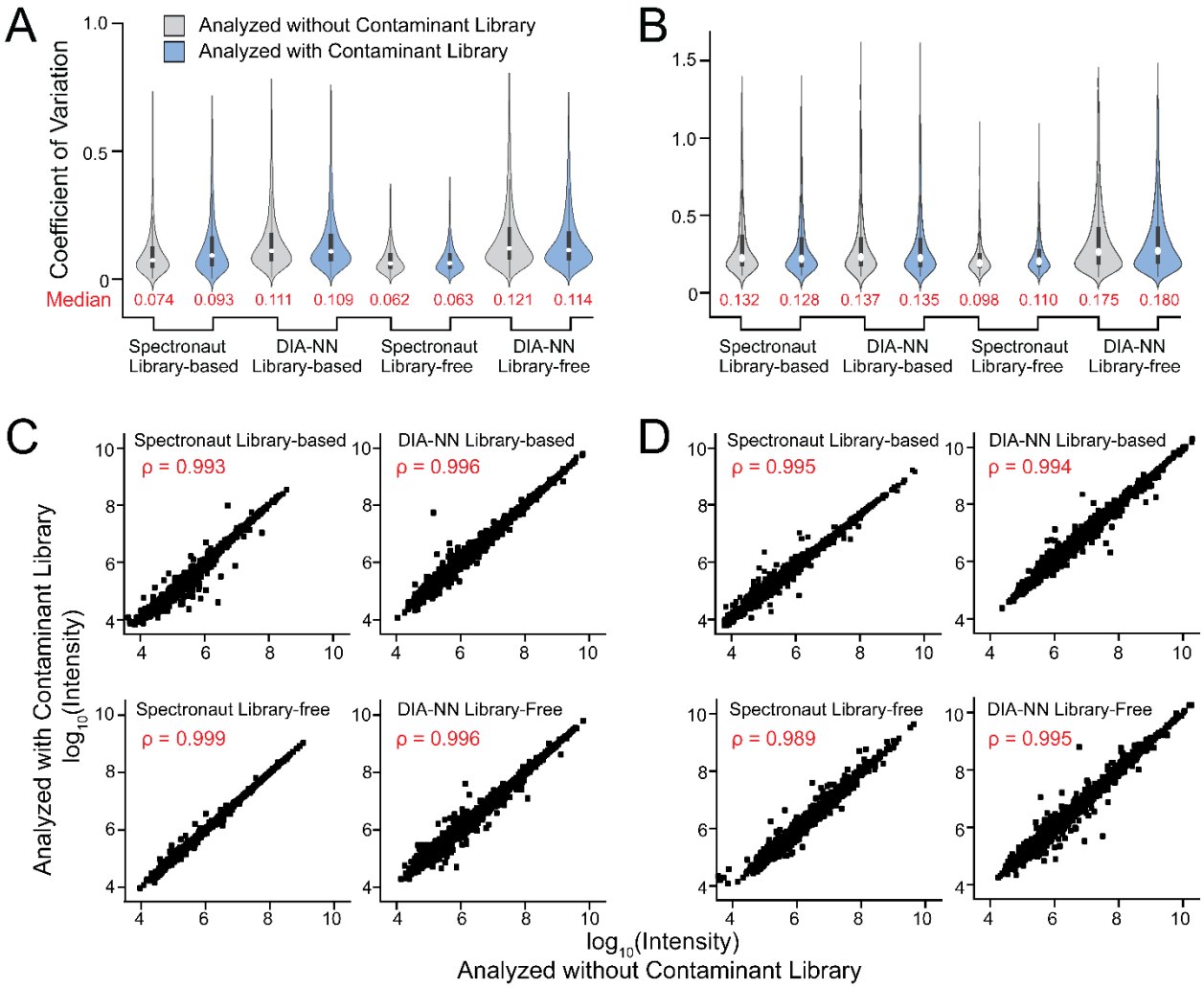

Figure 6: Evaluation of Protein Quantification Influenced by the Protein Contaminant Libraries. Violin boxplots showing the coefficient of variance for protein quantification with (blue) and without (grey) the contaminant library in HEK cells (A) and mouse brain tissue (B) DIA proteomics datasets. Spearman's correlation of protein intensities were calculated with and without the contaminant libraries in HEK cells (C) and mouse brain tissue (D).