Technical Brief

Improving protein detection confidence using SWATH mass spectrometry with large peptide reference libraries

Jemma X Wu[1], Dana Pascovici[1], Vera Ignjatovic[2], Xiaomin Song[1], Christoph Krisp[1], Mark P Molloy[1*]

[1]Australian Proteome Analysis Facility (APAF), Dept. Chemistry & Biomolecular Sciences, Macquarie University, Sydney, Australia

[2]Hematology Research Laboratory, Murdoch Childrens Research Institute, Melbourne, VIC 3052, Australia

*Corresponding author:

Assoc. Prof. Mark P. Molloy

Tel: +61 2 9850 6201

Fax: +61 2 9850 6200

mmolloy@proteome.org.au

Abstract

Protein quantification using data-independent acquisition methods such as SWATH-MS most commonly relies on spectral matching to a reference MS/MS assay library. To enable deep proteome coverage and efficient use of existing data, in-silico approaches have been described to use archived or publicly available large reference spectral libraries for spectral matching. Since implicit in the use of larger libraries is the increasing likelihood of false-discoveries, new workflows are needed to ensure high confidence in protein matching under these conditions. We present a workflow which introduces a range of filters and thresholds aimed at increasing confidence that the resulting proteins are reliably detected and their quantitation is consistent and reproducible. We demonstrated the workflow using extended libraries with SWATH data from human plasma samples and yeast-spiked human K562 cell lysate digest.

Keywords

SWATH, spectral library, false-discovery rate, quantification

There is increasing interest in the use of data-independent acquisition (DIA) strategies in proteomics[1, 2]. One major advantage is that more reliable run-to-run peptide quantitation can be achieved with DIA, which is difficult with information dependent acquisition (IDA) due to the stochastic nature of precursor ion selection and the complexity of biological samples [3, 4]. One of the more developed versions of DIA is known as SWATH[5], as implemented on Sciex Q-ToF mass spectrometers. Two main steps are involved in a typical SWATH experiment: (i) peptide MS/MS assay library generation and, (ii) SWATH data acquisition and peak area extraction. The peptide MS/MS assay library contains characteristic information about the peptide spectral features and retention times. MS/MS spectra generated by SWATH acquisition are interrogated against this library, allowing peptide quantification results to be extracted from ion chromatograms. High quality reference peptide MS/MS libraries are crucial to generating reliable SWATH quantitation. To enable deep SWATH data coverage, reference peptide libraries can be built by combining data from numerous IDA runs, but this is time consuming and can be technically challenging. For example, in plasma, which displays a large dynamic range of protein abundances, even extensive peptide fractionation results in only modest increases in reference library depth. An alternative strategy is to utilise the growing number of publicly available peptide reference libraries for matching or enhancing IDA-generated local reference libraries [6, 7].

We previously described an in-silico approach, *SwathXtend*, for extending the depth of in-house reference libraries by merging them with external or archived reference libraries without requiring iRT peptides for chromatographic alignment[8]. Other automated SWATH data analysis workflows which enable automatic cross-run alignment and peak picking exist but require synthetic iRT peptides to be spiked into every sample[9, 10]. *SwathXtend* can also work on libraries with iRT peptides; for these libraries, the final merged library will retain the iRT peptides and the iRT values will be unchanged. Currently, SwathXtend is only compatible with PeakView[11] or OpenSWATH[12] format libraries. To use SwathXtend with libraries from other softwares (e.g. Spectronaut [13]), it is first necessary to convert the files to a PeakView or OpenSWATH format. *SwathXtend* has great utility for generating large libraries, but as libraries grow, so does the risk of increasing false-positives[14]. In this technical brief we introduce a new range of filters and quality checks to scrutinize the extended libraries, ultimately improving the reliability for carrying out comparative proteomic analyses by SWATH using large reference spectral libraries.

The workflow described here is applicable to any SWATH experiment regardless of sample type and experimental design, however a particular focus of our group has been on the analysis of plasma, given the interests in plasma biomarkers for clinical applications[15-18]. Thus, for the purpose of this report we utilised data from a recently published SWATH analysis of human plasma [18]. This dataset consists of 40 plasma samples from healthy people in different age groups ranging from neonates to adults, with SWATH acquisition carried out with repeat technical injections for each sample.

SwathXtend requires a seed MS/MS spectral library for building the extended library. The seed library in this study was a spectral MS/MS library generated from four IDA runs of pooled samples representing the 40 plasma samples. To extend the library we used an in-house filtered version of a publicly available plasma spectral MS/MS library which was generated from analysis of a human twin population[16]. The in-house filtering included: peptides longer than 25 amino acids were removed and peptides were filtered by precursor mass/charge range 350 – 1500 Da. This add-on library from Liu et al. is referred to as the UK plasma library throughout this paper. The seed and add-on library contain 3584 and 35635 peptides (148 and 1721 proteins), respectively.

The current workflow's starting point as shown in Figure 1 used a local seed library obtained by IDA and under identical chromatographic conditions as the SWATH acquired data, and one or more existing archived or external spectral reference libraries. There are three major components in this workflow: (i) library filtering, generation and validation, (ii) SWATH data extraction with *PeakView* and (iii) SWATH results filtering and comparing. We elaborate on each component below.

The library generation and validation workflow (Figure 1A) includes the following steps:

1. Optionally, the add-on libraries can be filtered to improve the quality of the extended library. For example, one could remove peptides with long sequences (e.g. > 25 amino acids), modified peptides, or peptide fragments which have mass-charge ratio less than 350 m/z or greater than 1500 m/z.

2. Both the seed library and the optimised add-on libraries go through a cleaning process which removes peptides with low confidence (< 0.99) and low intensity (< 5).

3. The seed library and each add-on library will be pairwise checked for their matching quality. These checks include retention time (RT) correlation and relative ion intensity (RII) correlation for common peptides as previously described [8]. The cut-offs applied are R^2> 0.8 and ρ > 0.6.

4. If the above checking thresholds are met, the seed library and add-on libraries will be merged using *SwathXtend* to generate an extended peptide MS/MS reference library. To keep the SWATH extraction results consistent between local seed library and extended library, it is recommended not to remove modification and shared peptides. *PeakView* has features to do this if desired.

5. The protein and peptide overlaps are checked to ensure that the extended library includes all proteins and peptides in the cleaned seed library.

After the extended library was checked and found to pass the validation criteria, the SWATH raw data was extracted using both the seed and extended library with the same set of parameters and settings. These parameter settings, as listed in [8], include: the maximum number of peptide per protein as 100, the number of fragment ions per peptide as 6, peptide identification confidence as 99%, SWATH FDR for exported peak group detection as 1%, XIC RT window as 10 min and XIC mass window as 75ppm. Two additional parameters, the exclusion of modified and/or shared peptides,

need to be emphasized here. Both parameters are important but easily neglected which can lead to inconsistent SWATH results between the seed and extended datasets. In this study both parameters were set as excluded.

The extracted SWATH results are compared and checked using the steps as shown in Figure 1B.

1. Filter SWATH results using two criteria: number of FDR passes and number of peptides. *PeakView* exports a matrix of FDR values for each detected peptide in each sample, calculated using a decoy database strategy [5, 19]. A low FDR score means that the respective peptide was identified with high confidence in that particular sample; by default, *PeakView* retains all peptides that were identified with FDR score < 0.01 in at least one sample. The FDR criterion uses the peptide-level FDR score to only retain those peptides which have at least N (1≤N≤number of samples) samples that pass FDR score < 0.01, thus were identified with high confidence in several samples. The peptide criterion filters the proteins by the number of peptides identified for that respective protein. These two criteria can be applied together though care should be taken to avoid over-filtering. The filtered results will go through the checks described in steps 2-4 below.

2. Check the peptide and protein coverage of the two filtered results; the results extracted using the extended library should contain most of the results extracted with the seed library. As the filtering criteria cut-off values get stricter (for instance FDR pass number increases from 1 to more), we expect the quantification confidence to improve but without reducing the protein and peptide coverage achieved with the seed library.

3. Check the FDR distributions of the filtered results. The overall FDR distributions are plotted using 8 bins which include 0, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.8 and 1. The first bin, 0 to 0.01, is the default FDR threshold for *PeakView*. The other bins simply cover the remaining range of all possible FDR values. As the FDR pass cut-off increases for the extended library, the percentages of the lower-bin FDR (high confidence identifications) should increase while those of the higher-bin (low confidence identifications) should decrease.

4. Check the quantification consistency between the common proteins of the seed and extended library based results. We use two measurements to check the quantification consistency: the sample quantification correlation and the Coefficient of Variation (CV). Both measurements are computed for each protein common to the seed and extended results, by using the peak areas across all samples in the seed and extended library-based filtered results. The quantification correlation is the Spearman correlation of the peak area of the seed and the extended and similarly, the CV is calculated as the standard deviation divided by the mean of the two values.

Using the workflow described above, an extended library was generated by merging the plasma seed library and the UK plasma add-on library. Cleaning of the libraries was performed and no modification peptides or miss-cleavage peptides were removed. The matching quality of the seed and add-on library was R^2=0.98 and RII ρ=0.64. The extended library contains 37959 peptides and 1730 proteins. The numbers of proteins and peptides in the seed, UK plasma and final extended library are shown in Figure 2A.

The SWATH analysis result extracted using the extended library showed similar peptide and protein coverage with regards to that extracted with the seed library alone, with only one out of 147 proteins missed by the extended library extraction (Figure 2B). When the FDR filtering rules were applied, the filtered set of proteins in the extended SWATH results decreased as the FDR pass increased, especially at the first few FDR cut-offs (Figure 2C). We annotated the extracted proteins with the UniProtKB subcellular location controlled vocabulary [20] and calculated the proportion of "secreted" proteins in the total extracted proteins. For the SWATH results using the extended library, the percentage of the secreted proteins increased as the FDR pass cut-off increased (Figure 2C overlay). This confirms that confidence in the identifications of the extracted proteins increases as the FDR pass cut-off increases. The FDR bin distributions also show that the portion of high-confidence peptide identifications (e.g., FDR < 0.01, dark grey areas) increased as the FDR filtering becomes stricter (Figure 2D). The quantification consistency measured by the coefficient of variation (CV) of the same sample in the seed and extended SWATH results is shown in Figure 2E. The CV decreases as the FDR cut-off value increases, especially from 1 to 5. The median CV calculated for the same protein was 19% for five sample passing FDR < 0.01. The correlation was calculated using Spearman correlation due to the long-tailed peak area distribution.

Table 1 shows the comparison results of the seed and extended SWATH results under various filtering methods and cut-off values. Both FDR and number of peptides filtering methods can improve both the identification and quantification confidence of the extended SWATH results.

We applied the same workflow using the extensive human proteome assay library of ~10,000 proteins as a library for extension [3], to benchmark the human cell lysate-yeast datasets we previously reported [8]. The results show similar patterns to those described above for human plasma and can be found in the Supplemental material.

The workflow described in this paper presents a strategy for increasing confidence in protein identifications obtained in SWATH analysis when using large reference libraries. We show that by applying various FDR filters, this improves the identification and quantification confidence and consistency. Investigators need to strike the right balance with FDR filters to ensure high confidence in protein identifications while maximising peptide detection. This workflow can be applied in ways which suit each individual experiment; for instance rapid gains in quality are made when requiring proteins in the extended library to be extracted with high confidence in two or three samples, or with more than one peptide. The package including the data, source code and automated pipelines for this paper is available for public downloading from our FTP server (ftp://ftp.proteome.org.au/ReliableSwathManuscriptCode/). The updated *SwathXtend* package including the workflow functions is available from Bioconductor (currently available in *Development* version, release 3.6).

Acknowledgments
This work was conducted at the Australian Proteome Analysis Facility supported by the Australian Government's National Collaborative Research Infrastructure Scheme.

REFERENCES

[1] Sajic, T., Liu, Y., Aebersold, R., Using data-independent, high-resolution mass spectrometry in protein biomarker research: Perspectives and clinical applications. *PROTEOMICS-Clinical Applications* 2015, *9*, 307-321.
[2] Hu, A., Noble, W. S., Wolf-Yadlin, A., Technical advances in proteomics: new developments in data-independent acquisition. *F1000Research* 2016, *5*.
[3] Rosenberger, G., Koh, C. C., Guo, T., Röst, H. L.*, et al.*, A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Scientific Data* 2014, *1*.
[4] Selevsek, N., Chang, C. Y., Gillet, L. C., Navarro, P.*, et al.*, Reproducible and Consistent Quantification of the Saccharomyces cerevisiae Proteome by SWATH-mass spectrometry. *Molecular & cellular proteomics : MCP* 2015, *14*, 739-749.
[5] Gillet, L. C., Navarro, P., Tate, S., Röst, H.*, et al.*, Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics* 2012, *11*.
[6] Zi, J., Zhang, S., Zhou, R., Zhou, B.*, et al.*, Expansion of the Ion Library for Mining SWATH-MS Data through Fractionation Proteomics. *Analytical Chemistry* 2014, *86*, 7242-7246.
[7] Tsou, C.-C., Avtonomov, D., Larsen, B., Tucholska, M.*, et al.*, DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Meth* 2015, *12*, 258-264.
[8] Wu, J. X., Song, X., Pascovici, D., Zaw, T.*, et al.*, SWATH mass spectrometry performance using extended peptide MS/MS assay libraries. *Molecular & Cellular Proteomics* 2016, *15*, 2501-2514.
[9] Röst, H. L., Liu, Y., D'Agostino, G., Zanella, M.*, et al.*, TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nature methods* 2016.
[10] Röst, H. L., Aebersold, R., Schubert, O. T., Automated SWATH data analysis using targeted extraction of ion chromatograms. *bioRxiv* 2016, 044552.
[11] Schreiber, A., Cox, D., Using PeakView® Software with the XIC Manager for Screening and Identification with High Confidence based on High Resolution and Accurate Mass LC-MS/MS. *Application Note AB SCIEX* 2011, 2170811-2170803.
[12] Rost, H. L., Rosenberger, G., Navarro, P., Gillet, L.*, et al.*, OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature biotechnology* 2014, *32*, 219-223.
[13] Bernhardt, O. M., Selevsek, N., Gillet, L. C., Rinner, O.*, et al.*, Spectronaut: A fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data. *In: 60th ASMS Conference on Mass Spectrometry and Allied Topics* 2012.
[14] Muntel, J., Xuan, Y., Berger, S. T., Reiter, L.*, et al.*, Advancing Urinary Protein Biomarker Discovery by Data-Independent Acquisition on a Quadrupole-Orbitrap Mass Spectrometer. *Journal of proteome research* 2015, *14*, 4752-4762.

[15] Song, X., Amirkhani, A., Wu, J. X., Pascovici, D.*, et al.*, Analytical performance of nano-LC-SRM using nondepleted human plasma over an 18-month period. *Proteomics* 2016, *16*, 2118-2127.
[16] Liu, Y., Buil, A., Collins, B. C., Gillet, L. C.*, et al.*, Quantitative variability of 342 plasma proteins in a human twin population. *Molecular systems biology* 2015, *11*, 786.
[17] Song, X., Bandow, J., Sherman, J., Baker, J. D.*, et al.*, iTRAQ experimental design for plasma biomarker discovery. *Journal of proteome research* 2008, *7*, 2952-2958.
[18] Bjelosevic, S., Pascovici, D., Ping, H., Karlaftis, V.*, et al.*, Quantitative age-specific variability of plasma proteins in healthy neonates, children and adults. *Molecular & Cellular Proteomics* 2017.
[19] Reiter, L., Rinner, O., Picotti, P., Hüttenhain, R.*, et al.*, mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nature methods* 2011, *8*, 430-435.
[20] Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M.*, et al.*, UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Plant Bioinformatics: Methods and Protocols* 2016, 23-54.

Figure 1. Workflow for building reliable extended peptide MS/MS assay reference libraries. A) The extended library generation and validation; the red dashed square indicates functionality already present in the *SwathXtend* software [8]. B) The SWATH results filtering, checking and comparison of results.

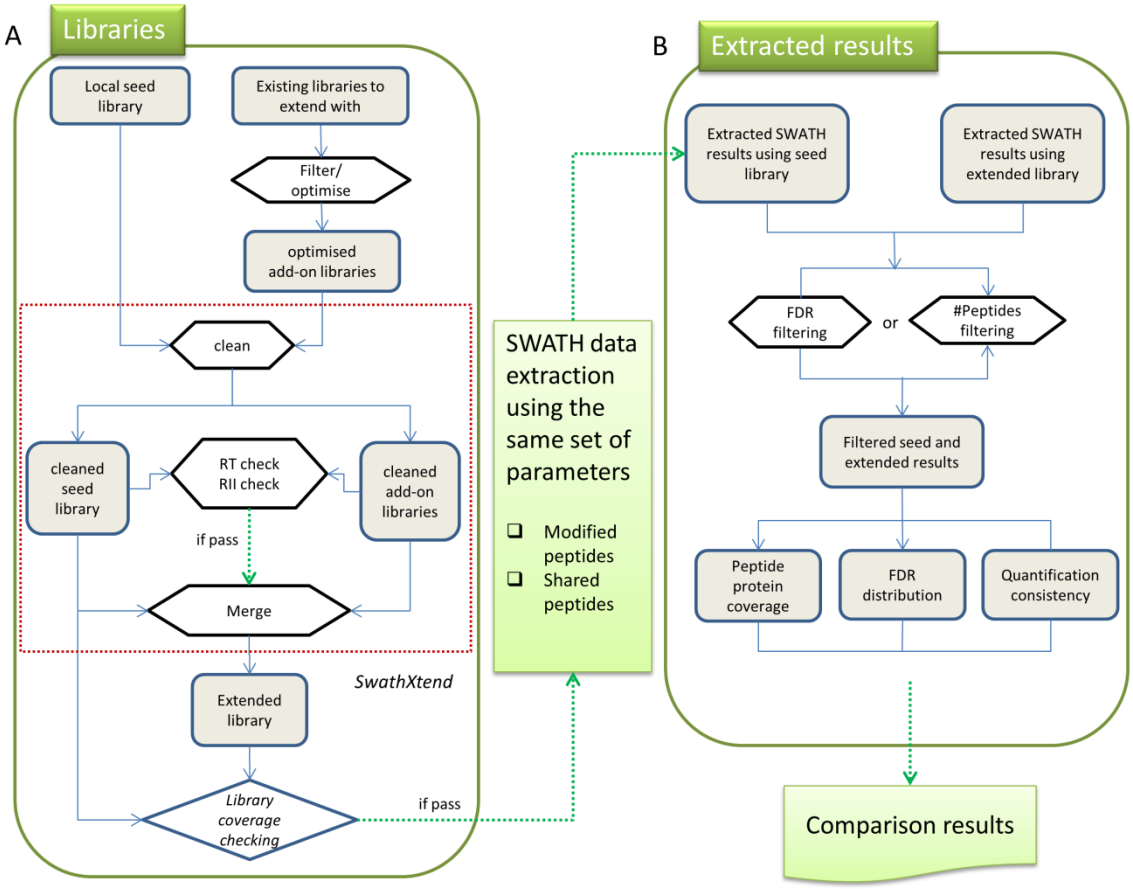

Figure 2. Results of seed and extended library for age-specific human plasma sample dataset. A) Number of peptides and proteins in various libraries; B) Venn diagrams of peptides and proteins in SWATH results extracted using the seed library and UK-plasma-extended library show a high overlap; C) Protein filtered by FDR with the percentage of secreted proteins overlayed shows that as stricter filtering is applied for the extended library the number of protein identifications decreases, but the percentage of secreted proteins in the extracted set increases; D) Percentage of FDR bins for different number of FDR passes shows that as stricter filtering is applied for the extended library the percentage of peptides identified with high confidence (dark grey) increases; E) Notched boxplots for protein quantification CV between the SWATH results extracted using seed and extended libraries improves with stricter filtering; the box contains the interquartile range (IQR), the whiskers (bars) extend to 1.5*IQR from the box.

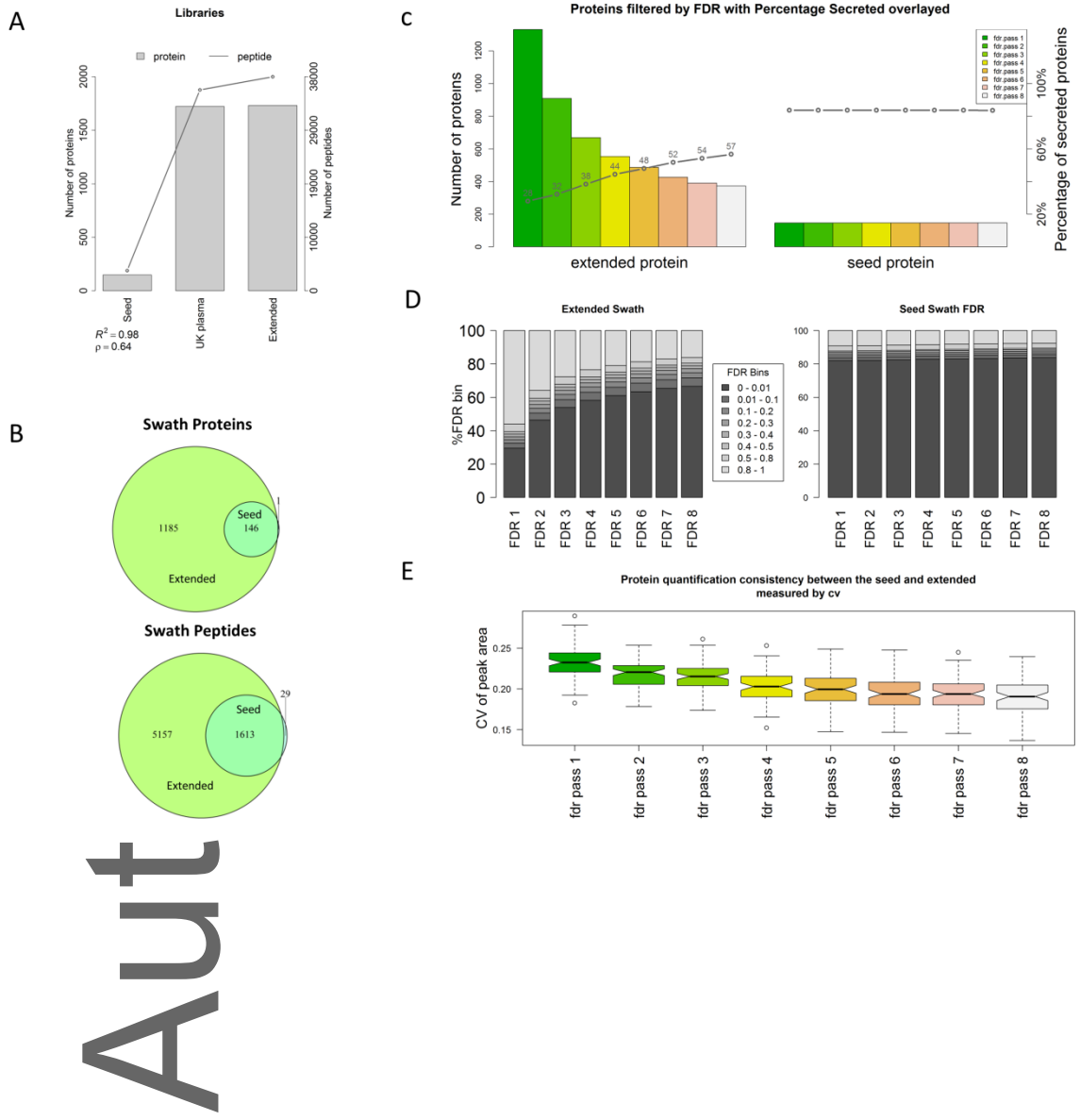

Table 1. Number of proteins, peptides and quantification consistency between seed and extended libraries using different filtering criteria. Note that stricter FDR filtering criterion does not reduce the number of proteins in the seed library, but reduces the number of proteins in the extended library.

	#Peptide			#Protein			Quantitation	
Filter applied	extended	seed	common	extended	seed	common	median correlation (Spearman)	median CV
Default (1 peptide, 1 sample pass FDR filter)	6972	1955	1839	1331	147	146	0.88	0.23
2 samples pass FDR filter	4397	1952	1812	909	147	145	0.88	0.22
3 samples pass FDR filter	3756	1944	1799	669	147	145	0.88	0.22
4 samples pass FDR filter	3464	1934	1778	553	147	144	0.89	0.2
5 samples pass FDR filter	3292	1929	1768	487	147	144	0.89	0.2
6 samples pass FDR filter	3153	1921	1753	427	147	144	0.89	0.19
7 samples pass FDR filter	3045	1915	1739	391	147	144	0.89	0.19
8 samples pass FDR filter	2987	1909	1731	372	146	143	0.88	0.19
#Peptide >= 2	6620	1931	1816	945	123	123	0.89	0.23
#Peptide >= 3	6116	1907	1791	676	110	109	0.88	0.24