Dataset Brief

Spectral libraries for SWATH-MS assays for *Drosophila melanogaster* and *Solanum lycopersicum*

Bertrand Fabre[1,2,4,7*], Dagmara Korona[3,4*], Clara I. Mata-Martinez[5], Harriet T. Parsons[1,2,4,6], Michael J. Deery[1,2,4], Maarten L. A. T. M. Hertog[5], Bart M. Nicolaï[5], Steven Russell[3,4], Kathryn S. Lilley[1,2,4]

1. Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, Cambridge, U.K

2. Department of Biochemistry, University of Cambridge, University of Cambridge, Cambridge, U.K

3. Department of Genetics, University of Cambridge, University of Cambridge, Cambridge, U.K

4. Cambridge Systems Biology Centre, University of Cambridge, Cambridge, U.K

5. BIOSYST-MeBioS, KU Leuven, Willem de Croylaan 42 - box 2428, 3001 Leuven, Belgium

6. Department of Plant and Environmental Sciences, Copenhagen University, Denmark

7. Current address: Bertrand Fabre, Technion Integrated Cancer Center (TICC), The Rappaport Faculty of Medicine and Research Institute, Haifa, Israel

* These authors contributed equally to this work.

Correspondence:

Pr. Kathryn S. Lilley, Ph.D., Cambridge Centre for Proteomics, Cambridge Systems Biology Centre, Department of Biochemistry, University of Cambridge, Cambridge CB2 1QR, U.K. Telephone: 01223 760255; e-mail: k.s.lilley@bioc.cam.ac.uk

Dr. Bertrand Fabre, Ph.D., Technion Integrated Cancer Center (TICC), The Rappaport Faculty of Medicine and Research Institute, Haifa, Israel, e-mail: bertrand.fabre@cantab.net

Abstract

Quantitative proteomics methods have emerged as powerful tools for measuring protein expression changes at the proteome level. Using mass-spectrometry (MS) based approaches, it is now possible to routinely quantify thousands of proteins. However, pre-fractionation of the samples at the protein or peptide level is usually necessary to go deep into the proteome, increasing both MS analysis time and technical variability. Recently, a new MS acquisition method named SWATH was introduced with the potential to provide good coverage of the proteome as well as a good measurement precision without prior sample fractionation. In contrast to shotgun based MS however, a library containing experimental acquired spectra is necessary for the bioinformatics analysis of SWATH data. In this study, we built spectral libraries for two widely used models to study crop ripening or animal embryogenesis, *Solanum lycopersicum* (tomato) and *Drosophila melanogaster*, respectively. The spectral libraries comprise fragments for 5,197 and 6,040 proteins for *Solanum lycopersicum* and *Drosophila melanogaster,* respectively, and allow reproducible quantification for thousands of peptides per MS analysis. The spectral libraries and all mass-spectrometry data are available in the MassIVE repository with the dataset identifiers

MSV000081074 and MSV000081075 and the PRIDE repository with the dataset identifiers PXD006493 and PXD006495.

Quantitative mass-spectrometry methods have been developed to monitor changes in proteins expression [1] or their subcellular localisation [2] but also protein-complexes composition [3], stoichiometry [4] and structural organisation [1]. Recently, a new approach based on Data Independent Acquisition (DIA) methods has been introduced which combines the extensive proteome coverage of DDA (Data Dependent Acquisition) approaches with the selectivity of targeted proteomics [5]. In this approach, called SWATH, precursors within defined m/z widows are co-fragmented, resulting in mixed MS/MS spectra. Extracted ion chromatograms (XICs) of the fragment are generated and the co-eluting peaks of the fragments of each precursor are used in the quantitative analysis. SWATH analysis is most commonly performed using a spectral library [6] although several approaches have been designed to identify peptides directly from the SWATH-MS files such as DIA-umpire developed by Nesvizhskii's lab [7] and Demux and Pecan developed by MacCoss's lab [8]. For DIA, the spectral library contains experimental acquired spectra of the precursors and can be produced directly from the SWATH data, by reconstituting the MSMS spectra using the retention time of the precursors and the fragments [7], or by DDA analysis of the sample, which can be fractionated to increase the number of spectra in the library [9]. Ideally, the spectral library should be generated on a MS instrument similar to the one used to acquire further SWATH-MS data as the correlation of the fragment intensities for a peptide acquired on different instrument was shown to be rather low [10]. Although Rosenberger et al. recently published a library containing assays for 10,000 human proteins [9] and deep libraries for other applications such as SRM (Selected Reaction Monitoring) have been generated [11-13], the number of publicly available spectral libraries for SWATH-MS is still limited. Several other studies

have produced libraries for other species [14-16]. However, spectral libraries are still missing for many species and, when available, the size of the libraries are limited and a deeper coverage of the proteome would increase the number of potential identifications in SWATH-MS analysis. In the present study, we produced spectral libraries for two well established models to study fruit ripening or animal embryogenesis, respectively *Solanum lycopersicum* L. and *Drosophila melanogaster*, for which no or low depth spectral libraries for SWATH-MS on TripleTOF instruments have been produced so far [17, 18]. These libraries contain assays for more than 5,000 proteins for both species and are best suited for the analysis of SWATH-MS data acquired on TripleTOF instruments. The precision and repeatability of the quantitative data generated from the analysis of SWATH data using these libraries were assessed using technical replicates for each species and compared to those obtained with other MS platforms.

To produce the spectral libraries, we first extracted proteins from Drosophila embryos and adult flies and tomato pericarps as previously described [17, 19]. The proteins were run on a 1D SDS-PAGE gel and digested with trypsin using in-gel digestion adapted from [20] and peptides were fractionated by high-pH reverse phase chromatography to increase the proteome coverage (Figure 1A and Supplementary information). HRM (Hyper Reaction Monitoring) peptides were spiked into the peptides mixtures and 10 µg of each fraction was injected to a Sciex TripleTOF 6600 mass spectrometer (Sciex, Framingham, MA, USA) fitted with a microflow set-up as described in [17] (Figure 1A). HRM peptides allow precise retention time alignment and their addition in the samples for the generation of the spectral library as well as the SWATH acquired samples is necessary to analyse the SWATH-MS data using software such as Spectronaut [21]. The resulting .wiff files (60 files for tomato and 72 files for Drosophila) were analysed using MaxQuant (version 1.5.4.3) [22] and Spectronaut (version 10) [21] (Supporting Information) (Figure 1A). Peptide minimal length was set to 7 amino acids and up to 2 missed cleavage were allowed. A

false discovery rate (FDR) of 1% was applied both at the peptide and protein levels in MaxQuant. Protein interference was performed in Spectronaut and peptides with less than 3 fragments were discarded. The best 3 to 6 fragments per precursor were conserved, the majority of the precursor having 6 fragments (Figure S1). Using these parameters, the spectral libraries for Drosophila and tomato contain 277,238 and 162,882 assays (fragments), respectively (Figure 1B). These fragments match to 47,810 and 28,516 precursors from 6,040 and 5,197 protein groups for Drosophila and tomato respectively (Figure 1B). For Drosophila melanogaster, adding a fractionated adult sample to the spectral library increased the number of proteins in the library from 5,335 to 6,040 showing that adding specific tissue/developmental stage is useful to increase the size of spectral libraries for SWATH-MS. The peptide size distributions are very similar for both spectral libraries with a high proportion of peptides around 10 amino acids long (Figure S2) which is in agreement with previous study using trypsin in yeast [23]. Our spectral libraries cover around 34% and 15% of the expected proteomes of *Drosophila melanogaster* and *Solanum lycopersicum* which, to our knowledge, constitute the most complete spectral libraries acquired on TripleTOF instruments for either species.

In order to test the suitability of the spectral libraries generated for SWATH-MS data analysis, we measured the repeatability and precision of the SWATH-MS assays using our spectral libraries by performing four injection replicates of 5 µg of a Drosophila embryo sample and a membrane preparation from tomato pericarp using a SWATH acquisition mode as previously described [17]. SWATH data were analysed using Spectronaut. When selecting a q-value of 0.01 (FDR of 1%), on average 3,444 peptides from the membrane preparation of tomato pericarp and 9,180 peptides from the Drosophila embryo sample were quantified with good reproducibility as shown by median coefficients of variation of 5.3% for tomato and 7.5% for Drosophila (Figure 2A and 2B). When changing the q-value, we observed an increase in the number of peptides identified with a

correlated increase of the coefficient of variation (Figure 2A and 2B and Figure S3). However, even with a FDR of 5%, we found the technical reproducibility to be similar (median CVs below 8.5% for Drosophila and tomato) to other studies [9, 24].

Next, we assessed the performance of SWATH-MS analysed with our spectral libraries compared to other DDA mode on the same or different MS platforms. For this comparison, a Drosophila embryo sample was injected four times on a Q Exactive (Fisher Scientific, Waltham, MA, USA), a LTQ-Orbitrap Velos (Fisher Scientific, Waltham, MA, USA) or a TripleTOF 6600 mass spectrometer, all in DDA acquisition mode, and on a TripleTOF 6600 in SWATH acquisition mode (Supporting Information). DDA data from the Q Exactive, LTQ-Orbitrap Velos and TripleTOF 6600 were analysed with MaxQuant (Supporting Information) with FDRs of 1% at peptide and protein levels. The SWATH data were analysed with Spectronaut with a q-value of 0.01 (Supporting Information). In SWATH acquisition mode the TripleTOF 6600 provided on average 77.8% more identifications than in DDA acquisition mode and 56.4% more identifications than the LTQ-Orbitrap Velos (Figure 2C). The Q Exactive identified, on average, only 2.9% more peptides than the TripleTOF 6600 in SWATH acquisition mode (Figure 2C). When comparing pairwise the measured peptides intensities between injection replicates, we observed that the TripleTOF 6600 in SWATH acquisition mode provides very reproducible measurements compared to other MS platforms as shown by the Fisher z transformed Pearson correlation coefficients above 2.65 (corresponding to $r = 0.99$, Figure 2D and Figure S4). Moreover, the coefficient of variations of the measured peptides intensities between injection replicates were lower for the TripleTOF 6600 in SWATH acquisition mode compared to the other MS platforms (Figure 2E). For all MS platforms, low intense peptides display higher signal variability than higher intense ones (Figure S5). Coefficient of variations of less than 20% were observed for 94.6% of the measured peptides with the TripleTOF 6600 in SWATH acquisition mode (Figure 2E). These results demonstrate the suitability

of the spectral libraries generated in the present study to provide reproducible quantitative measurements for the analysis of SWATH-MS data.

In summary we provide through the present study a set of two spectral libraries for *Solanum lycopersicum* and *Drosophila melanogaster* to analyse SWATH-MS data. We show that SWATH-MS using these spectral libraries allows reproducible quantification for thousands of peptides in tomato or Drosophila samples. These libraries, and all raw files, were uploaded in the MassIVE repository with the dataset identifier MSV000081074 and MSV000081075 and the PRIDE repository [25] with the dataset identifier PXD006493 and PXD006495 and can be used with software like OpenMS and Spectronaut. The raw data can also serve as a foundation for the generation of extended (with data acquired on similar or different MS platforms) or customized (e.g. for post-translational modifications) spectral libraries to explore the biology of *Solanum lycopersicum* and *Drosophila melanogaster*.

Acknowledgements

B.F. is funded by Biotechnology and Biological Science Research Council (Ref: BB/L002817/1) and a long term EMBO fellow (ALTF 1204-2015) cofounded by Marie Curie Actions (LTFCOFUND2013, GA-2013-609409). D.K. is funded by Biotechnology and Biological Science Research Council (Ref: BB/L002817/1). This work was partially supported by the Bilateral Scientific Research Cooperation Projects FWO.106.2013.20 between NAFOSTED (Vietnam) and FWO Flanders. We want to thank the Cost Action FA1106 QualityFruit to provide C.I.M. with a Short Term Scientific Mission to carry out her research in the Cambridge Centre for Proteomics.

The authors have declared no conflict of interest.

References

[1] Aebersold, R., Mann, M., Mass-spectrometric exploration of proteome structure and function. *Nature* 2016, *537*, 347-355.

[2] Christoforou, A., Mulvey, C. M., Breckels, L. M., Geladaki, A.*, et al.*, A draft map of the mouse pluripotent stem cell spatial proteome. *Nature communications* 2016, *7*, 8992.

[3] Olshina, M. A., Sharon, M., Mass Spectrometry: A Technique of Many Faces. *Quarterly reviews of biophysics* 2016, *49*.

[4] Liko, I., Allison, T. M., Hopper, J. T., Robinson, C. V., Mass spectrometry guided structural biology. *Current opinion in structural biology* 2016, *40*, 136-144.

[5] Gillet, L. C., Navarro, P., Tate, S., Rost, 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 : MCP* 2012, *11*, O111 016717.

[6] Anjo, S. I., Santa, C., Manadas, B., SWATH-MS as a tool for biomarker discovery: From basic research to clinical applications. *Proteomics* 2017, *17*.

[7] Tsou, C. C., Avtonomov, D., Larsen, B., Tucholska, M.*, et al.*, DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature methods* 2015, *12*, 258-264, 257 p following 264.

[8] Bern, M., Finney, G., Hoopmann, M. R., Merrihew, G.*, et al.*, Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Analytical chemistry* 2010, *82*, 833-841.

[9] Rosenberger, G., Koh, C. C., Guo, T., Rost, H. L.*, et al.*, A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Scientific data* 2014, *1*, 140031.

[10] 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 : MCP* 2016, *15*, 2501-2514.

[11] Brunner, E., Ahrens, C. H., Mohanty, S., Baetschmann, H.*, et al.*, A high-quality catalog of the Drosophila melanogaster proteome. *Nature biotechnology* 2007, *25*, 576-583.

[12] Kusebauch, U., Campbell, D. S., Deutsch, E. W., Chu, C. S.*, et al.*, Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome. *Cell* 2016, *166*, 766-778.

[13] Loevenich, S. N., Brunner, E., King, N. L., Deutsch, E. W.*, et al.*, The Drosophila melanogaster PeptideAtlas facilitates the use of peptide data for improved fly proteomics and genome annotation. *BMC bioinformatics* 2009, *10*, 59.

[14] Malmstrom, E., Kilsgard, O., Hauri, S., Smeds, E.*, et al.*, Large-scale inference of protein tissue origin in gram-positive sepsis plasma using quantitative targeted proteomics. *Nature communications* 2016, *7*, 10261.

[15] Muller, D. B., Schubert, O. T., Rost, H., Aebersold, R., Vorholt, J. A., Systems-level Proteomics of Two Ubiquitous Leaf Commensals Reveals Complementary Adaptive Traits for Phyllosphere Colonization. *Molecular & cellular proteomics : MCP* 2016, *15*, 3256-3269.

[16] Schubert, O. T., Ludwig, C., Kogadeeva, M., Zimmermann, M.*, et al.*, Absolute Proteome Composition and Dynamics during Dormancy and Resuscitation of Mycobacterium tuberculosis. *Cell host & microbe* 2015, *18*, 96-108.

[17] Fabre, B., Korona, D., Groen, A., Vowinckel, J.*, et al.*, Analysis of Drosophila melanogaster proteome dynamics during embryonic development by a combination of label-free proteomics approaches. *Proteomics* 2016, *16*, 2068-2080.

[18] Okada, H., Ebhardt, H. A., Vonesch, S. C., Aebersold, R., Hafen, E., Proteome-wide association studies identify biochemical modules associated with a wing-size phenotype in Drosophila melanogaster. *Nature communications* 2016, *7*, 12649.

[19] Mata, C. I., Fabre, B., Hertog, M. L., Parsons, H. T.*, et al.*, In-depth characterization of the tomato fruit pericarp proteome. *Proteomics* 2017, *17*.
[20] Fabre, B., Lambour, T., Garrigues, L., Amalric, F.*, et al.*, Deciphering preferential interactions within supramolecular protein complexes: the proteasome case. *Molecular systems biology* 2015, *11*, 771.
[21] Bruderer, R., Bernhardt, O. M., Gandhi, T., Miladinovic, S. M.*, et al.*, Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Molecular & cellular proteomics : MCP* 2015, *14*, 1400-1410.
[22] Cox, J., Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* 2008, *26*, 1367-1372.
[23] Swaney, D. L., Wenger, C. D., Coon, J. J., Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *Journal of proteome research* 2010, *9*, 1323-1329.
[24] 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.
[25] Vizcaino, J. A., Deutsch, E. W., Wang, R., Csordas, A.*, et al.*, ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology* 2014, *32*, 223-226.

Figure 1: Workflow used in this study to produce *Drosophila melanogaster* and *Solanum Lycopersicum* spectral libraries

A. Embryos/adult flies and pericarps, respectively from *Drosophila melanogaster* and *Solanum lycopersicum* were lysed and the proteins were digested in gel with trypsin. The peptides were fractionated by high pH reverse phase chromatography, HRM peptides were spiked and the samples were injected on a Sciex Triple TOF 6600. MaxQuant was used to analyse the .wiff files and Spectronaut was used to generate the libraries. B. Numbers of fragments, precursors and protein groups present in the spectral libraries.

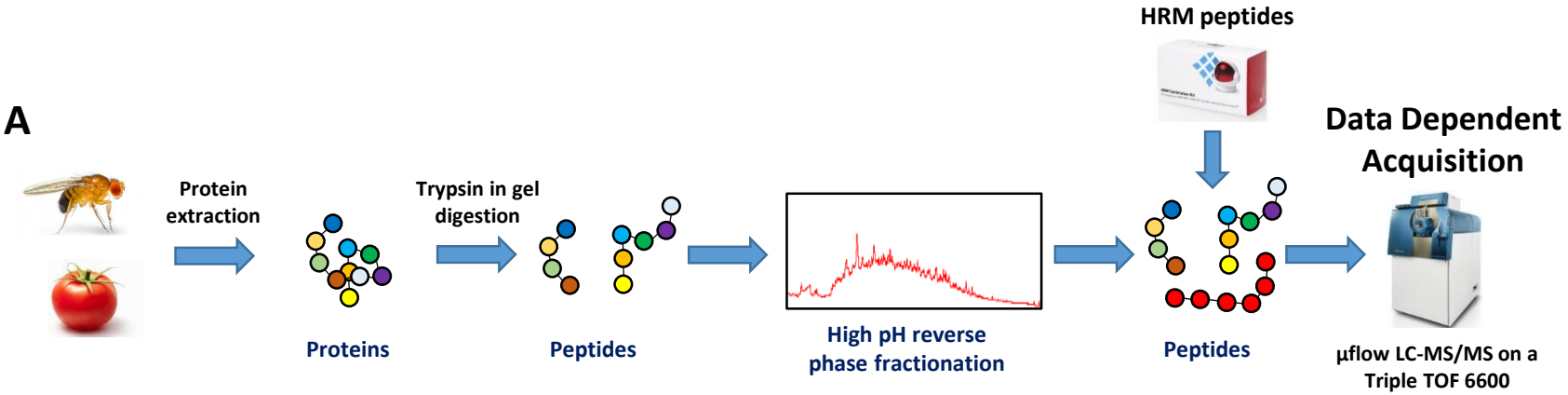
Figure 2: Number of peptide identifications and reproducibility of SWATH-MS runs using the *Drosophila melanogaster* and *Solanum Lycopersicum* spectral libraries

A. and B. Number of peptides identified and coefficients of variation measured from 4 injection replicates of a membrane samples of *Solanum Lycopersicum* (A) and an embryo sample of *Drosophila melanogaster* (B) in SWATH acquisition mode. The black bars represent identified peptides numbers and the lines the coefficient of variation (dashed lines represents Q3 and Q1 CVs and the solid line represents median CV). The results are shown for several FDR values. Errors bars represent standard deviations. C. Number of peptides identified with a FDR of 1% from 4 injection replicates of an embryo sample of *Drosophila melanogaster* on different MS platforms (or acquisition modes). The black bars represent identified peptides numbers. Errors bars represent standard deviations. D. Pairwise comparison of peptides intensities measured from 4 injection replicates of an embryo sample of *Drosophila melanogaster* on different MS platforms (or acquisition modes). The dots represent the Fisher Z transformed Pearson correlation coefficients resulting from the pairwise comparison. E. Box plots of the coefficients of variation

measured from 4 injection replicates of an embryo sample of *Drosophila melanogaster* on different MS platforms (or acquisition modes).

Figure 1

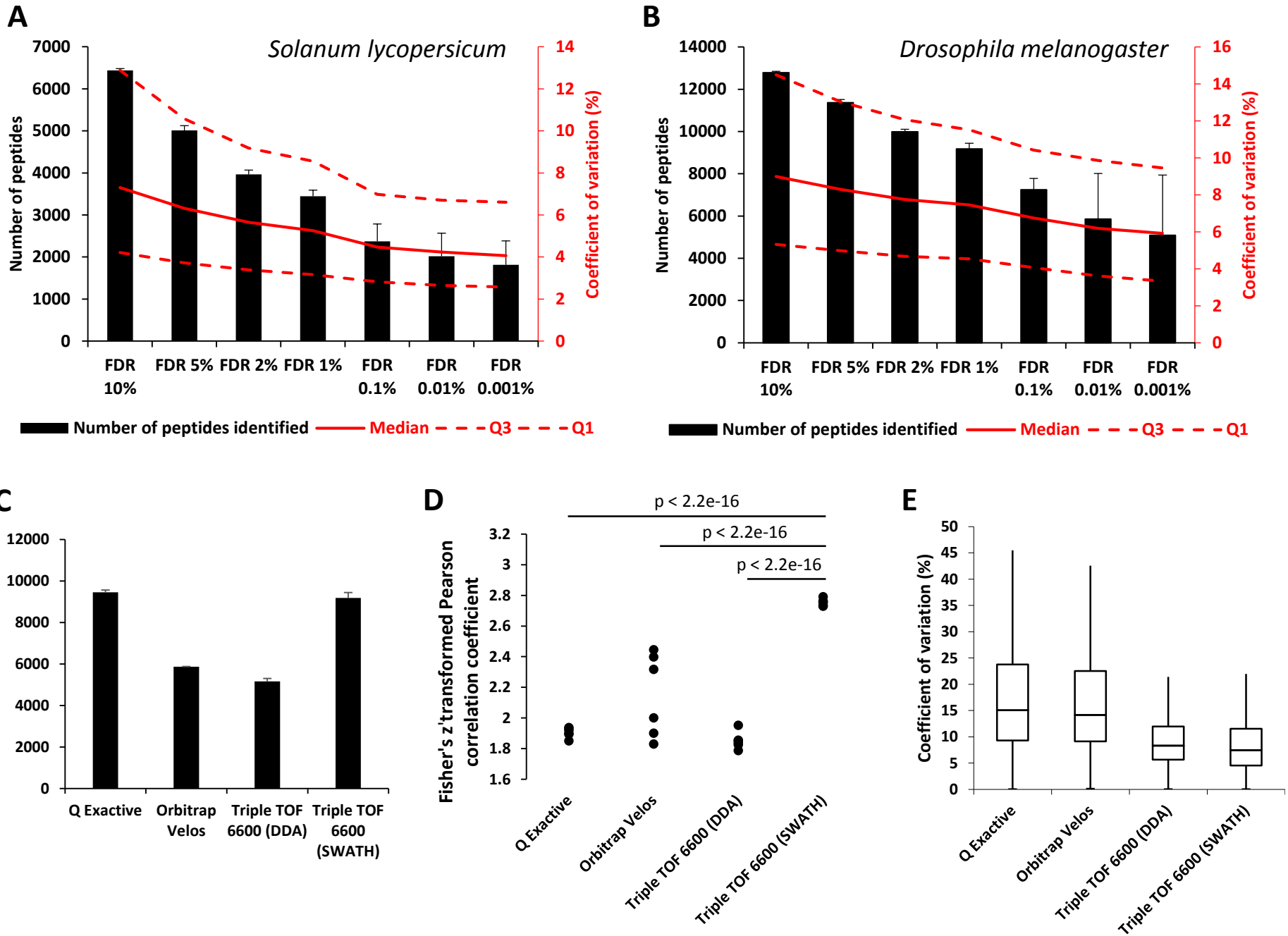
A



B

	Fragments	Precursors	Protein Groups
Drosophila melanogaster	277,238	47,810	6,040
Solanum Lycopersicum	162,882	28,516	5,197

Figure 2


A
Solanum lycopersicum
Number of peptides
Coefficient of variation (%)
FDR 10%
FDR 5%
FDR 2%
FDR 1%
FDR 0.1%
FDR 0.01%
FDR 0.001%
Number of peptides identified
Median
Q3
Q1
B
Drosophila melanogaster
C
Number of peptides identified
Q Exactive
Orbitrap Velos
Triple TOF 6600 (DDA)
Triple TOF 6600 (SWATH)
D
p < 2.2e-16
p < 2.2e-16
p < 2.2e-16
Fisher's z'transformed Pearson correlation coefficient
E
Coefficient of variation (%)