Hybrid spectral library combining DIA-MS data and a targeted virtual library substantially deepens the proteome coverage

Ronghui Lou, Pan Tang, Kang Ding, Shanshan Li, Cuiping Tian, Yunxia Li, Suwen Zhao, Yaoyang Zhang, Wenqing Shui

PII: S2589-0042(20)30087-0
DOI: https://doi.org/10.1016/j.isci.2020.100903
Reference: ISCI 100903

To appear in: ISCIENCE

Received Date: 9 December 2019
Revised Date: 17 January 2020
Accepted Date: 6 February 2020

Please cite this article as: Lou, R., Tang, P., Ding, K., Li, S., Tian, C., Li, Y., Zhao, S., Zhang, Y., Shui, W., Hybrid spectral library combining DIA-MS data and a targeted virtual library substantially deepens the proteome coverage, ISCIENCE (2020), doi: https://doi.org/10.1016/j.isci.2020.100903.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2020
Hybrid spectral library combining DIA-MS data and a targeted virtual library substantially deepens the proteome coverage

Ronghui Lou\textsuperscript{1,2,3‡}, Pan Tang\textsuperscript{1,2,3‡}, Kang Ding\textsuperscript{1,2,3‡}, Shanshan Li\textsuperscript{1}, Cuiping Tian\textsuperscript{1}, Yunxia Li\textsuperscript{4}, Suwen Zhao\textsuperscript{1,*}, Yaoyang Zhang\textsuperscript{4,*}, Wenqing Shui\textsuperscript{1,2,5*}

\textsuperscript{1}iHuman Institute, ShanghaiTech University, Shanghai 201210, China
\textsuperscript{2}School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China
\textsuperscript{3}University of Chinese Academy of Sciences, Beijing 100049, China
\textsuperscript{4}Interdisciplinary Research Center on Biology and Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, 201210, China
\textsuperscript{5}Lead contact

\textsuperscript{‡}Equal contribution

\textsuperscript{*To whom correspondence should be addressed to:
Wenqing Shui Email: shuiwq@shanghaitech.edu.cn
Yaoyang Zhang Email: zyy@sioc.ac.cn
Suwen Zhao Email: zhaosw@shanghaitech.edu.cn
Abstract

Data-independent acquisition mass spectrometry (DIA-MS) is a powerful technique that enables relatively deep proteomic profiling with superior quantification reproducibility. DIA data mining predominantly relies on a spectral library of sufficient proteome coverage that, in most cases, is built on data-dependent acquisition-based analysis of the same sample. To expand the proteome coverage for a pre-determined protein family, we report herein on the construction of a hybrid spectral library that supplements a DIA experiment-derived library with a protein family-targeted virtual library predicted by deep learning. Leveraging this DIA hybrid library substantially deepens the coverage of three transmembrane protein families (G protein coupled receptors; ion channels; and transporters) in mouse brain tissues with increases in protein identification of 37-87%, and peptide identification of 58-161%. Moreover, of the 412 novel GPCR peptides exclusively identified with the DIA hybrid library strategy, 53.6% were validated as present in mouse brain tissues based on orthogonal experimental measurement.
Introduction

Data-independent acquisition mass spectrometry (DIA-MS) is emerging as a powerful technology for proteomics research owing to its superior accuracy and reproducibility in proteomic quantification while retaining relatively deep coverage of the proteome (Gillet et al., 2012; Ludwig et al., 2018). To maximize the proteome coverage in DIA-MS analyses, sample-specific spectral libraries are typically built based on peptide identifications from a conventional data-dependent acquisition (DDA) experiment, which often involve offline pre-fractionation of peptide samples (Schubert et al., 2015). Raw DIA data can then be processed for peptide identification and quantification using a peptide-centric scoring algorithm (Ting et al., 2015) against this DDA experiment-derived spectral library. Alternatively, DIA data can be processed and searched directly against a FASTA sequence database (Ting et al., 2017; Tsou et al., 2015), however, such library-free approaches usually result in lower proteome coverage (Ting et al., 2017).

Recent innovative work by Gessulat et al. (Gessulat et al., 2019) and Tiwary et al. (Tiwary et al., 2019) demonstrated the feasibility of building a virtual spectral library based on separate predictions of fragment ion intensities and peptide retention times from deep learning models. Fundamentally, these major proof-of-concept studies demonstrated that DDA experiment-derived spectral libraries can be replaced with virtual spectral libraries built for experimentally-detected peptides to achieve nearly equivalent whole-proteome coverage (Gessulat et al., 2019; Tiwary et al., 2019).

In theory, current deep learning models can make predictions for all peptides yielded from \textit{in silico} digestion of the entire proteome, but this would result in the substantial expansion of the spectral library and attendant significant increase of the false discovery rate (FDR) (Rosenberger et al., 2017). Given that many biological studies focus on a specific class of proteins, we propose a new strategy of constructing a targeted virtual library for a given protein superfamily to deepen its proteome coverage. Targeted MS assays have been developed for
the detection and quantification of a predetermined set of proteins in complex matrices across multiple samples (Picotti and Aebersold, 2012). However, these conventional assays have several drawbacks. They require tremendous initial effort to select optimal peptides and establish assay parameters for each given protein and instrument platform. Also, only tens of proteins are routinely measured in a single run with targeted MS assays (Kusebauch et al., 2016). Herein, we present an approach of exploiting a targeted virtual library to profile the expression of hundreds of proteins within a superfamily by single-injection DIA analysis while strictly controlling the FDR in data mining.

Results

Generating an initial DIA spectral library for mouse brain tissues

To evaluate our strategy, we chose transmembrane protein families as our targets. These proteins were selected because of their strong hydrophobicity, relative low abundance, and fast turn-over, which make them challenging to profile using conventional proteomics techniques. Specifically, we focused on the G protein-coupled receptor (GPCR) superfamily: the mouse genome contains 524 annotated GPCRs, each having seven transmembrane domains (Katritch et al., 2013). Given that GPCRs represent one of the most prominent classes of drug targets, particularly for neurological diseases (Huang et al., 2017; Katritch et al., 2013), the ability to deeply and accurately profile the GPCR sub-proteome will greatly benefit both basic neuroscience and therapeutic development. However, GPCRs are notoriously under-represented in proteomic datasets. A meta-proteome analysis of diverse human samples identified only 65 GPCR proteins (Wilhelm et al., 2014) of the 831 encoded in the human genome, a much lower proportion than for most other protein superfamilies. A similarly low identification rate for GPCRs was also reported in a very deep proteomic survey of human cells (56 GPCRs identified) (Bekker-Jensen et al., 2017) (Figure S1).
Our study examined the GPCR sub-proteome in three mouse brain tissues: cerebellum, midbrain, and spinal cord. We prepared cell membrane fractions and performed single-injection DIA-MS analysis of digested protein extracts for each tissue, and identified 36, 38, and 38 GPCRs based on spectral matches for 405, 402, and 381 peptide precursors in the three brain regions, respectively (Table S1). In our workflow, DIA data from 12 total analyses (from experimental quadruplets of each region) were directly searched against a mouse FASTA database using Spectronaut (Bruderer et al., 2015) to generate an initial spectral library (i.e., without building a sample-specific spectral library from DDA experiments; Fig. 1). This DIA spectral library for the brain samples included 415 peptide precursors mapped to 38 GPCR proteins.

**Constructing a GPCR-targeted virtual library with re-trained pDeep and DeepRT**

Before constructing a virtual spectral library, we tested the performance of several deep learning models to predict fragment ion intensities and retention time indices (iRT) for the 415 GPCR peptide precursors from the initial DIA spectral library. Distinct from the aforementioned whole-proteome virtual library approaches, we here used the deep neutral network-based models pDeep (Zhou et al., 2017) to predict fragment ion intensities and DeepRT (Ma et al., 2018) to predict iRT from GPCR peptide sequences (Fig. 1). The pre-trained pDeep model achieved excellent overall agreement between the experimental and predicted MSMS spectra at a defined collision energy for this GPCR peptide test set (median Pearson correlation coefficient = 0.93, median spectral angle = 0.85) (Figures S2A, S2B). However, for iRT prediction and because of large differences in the liquid chromatography conditions adopted in our DIA experiment vs. the earlier experiments upon which the DeepRT was originally trained (Ma et al., 2018), the initial performance of DeepRT was lower than expected (Figure S2C).

Unsatisfied with these pre-trained models, we next applied a transfer learning technique (Pan and Yang, 2010) to further train both the pDeep and DeepRT models using the majority of
non-GPCR peptide entries in our DIA spectral library (27,390 in total). Subsequent analysis of our GPCR peptide test set showed significant improvements compared to the original pre-trained models (e.g., ΔIRT\textsubscript{95} value of 38 units for the pre-trained DeepRT vs. 12.9 units for the re-trained model; Figure S2). Notably, when we evaluated another deep learning model (Prosit) with our GPCR peptide test set, we found that the prediction results for both fragment ion intensities and iRTs were no better or slightly worse than those obtained with our re-trained pDeep and DeepRT models (Figure S3).

Having obtained sufficient prediction performance with the two deep learning models, we next set out to construct a GPCR-targeted virtual library. Specifically, we used the re-trained models to predict fragment ion intensities and iRT values for tryptic peptides yielded from an \textit{in silico} digestion of the full complement of 524 GPCR proteins in the mouse genome (Fig. 1). To account for the known influence of digestion settings on the library size, we performed \textit{in silico} digestion with 12 different combinations of peptide charge state, numbers of missed cleavages, and Met oxidation status (referred to as P1–P12; Fig. 2A).

**Increasing the depth of GPCR identification by DIA-MS with a hybrid library**

Each GPCR virtual library generated with a given set of conditions was then merged with the experimental DIA spectral library, yielding 12 distinct hybrid libraries. Whereas the initial DIA spectral library comprised 34,922 precursors, the size of the 12 hybrid libraries ranged between 50,927–195,730 precursors (Fig. 2B, Table S2). Remarkably, searching our DIA data against each of the 12 hybrid libraries invariably led to drastically more putative GPCR identifications than searching with the initial DIA library. Taking the cerebellum sample as an example, an average of 114 GPCR proteins (from an average of 661 peptides) were putatively identified using the 12 virtual libraries, whereas only 36 proteins (from 304 peptides) were identified with the initial DIA library (Figs. 2C-D). In the best-performing library (P2), 737 peptides were
mapped to 136 GPCRs, representing a gain of 445 peptides and 102 GPCRs relative to the initial DIA library. Of note, this prominent increase in coverage was only observed for GPCRs; that is, there was no significant change in the total numbers of protein identifications when using the 12 hybrid libraries (Figure S4, Table S3). Moreover, similar increases in the numbers of putative GPCR peptides and proteins were also obtained when analyzing DIA data from the two other mouse brain regions (Figure S5).

**FDR assessment in use of the GPCR hybrid library**

The huge gains in the number of putatively identified GPCRs prompted us to closely examine the potential for false positives since the expanded library size from the incorporation of a targeted virtual library was expected to increase the error rate (Rosenberger et al., 2017). We first performed an additional DDA experiment by fractionating the peptide samples from each brain region (Table S4). Our concatenation of the GPCR identification lists from the initial DIA and the new DDA experiments yielded a "max identification (ID) list" (i.e., the maximal collection of experimentally identified GPCR proteins and peptides). We then calculated the percentages of *bona fide* identifications from these max ID lists that were recovered using our 12 hybrid libraries. The P2 hybrid library yielded the highest recovery rates: 92.5% of proteins and 75.6% of peptides in the cerebellum sample (Figs. 2C-D), with similar results for the other two brain regions (Figure S5). Thus, P2 was chosen as the best-performing condition for virtual library construction (missed cleavages 0/1, charge state 2/3, Met oxidation 0). Furthermore, we observed no loss in the reproducibility of quantification for total peptides or for GPCR peptides identified with the P2 hybrid library compared to the initial DIA spectral library (median CVs at 8.90%-10.50%; Figure S6).

Beyond these *bona fide* identifications, searching with the P2 hybrid library putatively identified 372 novel GPCR peptides representing 87 novel GPCRs, which were not present in the cerebellum max ID list (Figure S7). To estimate the error rate amongst these P2-hybrid-
library-only GPCR identifications, we created a decoy virtual library via the \textit{in silico} digestion of 524 reverse GPCR sequences (again using the P2 condition). After combining this decoy virtual library with the initial DIA spectral library, a DIA data search with this hybrid library yielded 127 decoy peptides and 77 decoy proteins in the cerebellum (Figs. 2E-F; Table S5). This result strongly indicated a significant FDR when using the default settings in Spectronaut, although we set FDR to <1% at the peptide and protein levels in the DIA data search. When we subsequently applied a more stringent data filtration cut-off (Cscore >0.9), 81.9% of the decoy peptides and 81.8% of the decoy proteins were removed, and the same data filtration retained 625 (84.8%) GPCR peptides and 61 (44.9%) GPCRs identified in the cerebellum with the P2 hybrid library (Figs. 2E-F). The other two regions showed similar results before and after data filtration (Figure S8, Table S5).

 Altogether 71 GPCR proteins and 810 GPCR peptides were identified from three mouse brain regions after applying the data filtration cut-off, compared to 38 GPCR proteins and 310 peptides identified with the initial DIA library (Figs. 4A, 4D). Thus, with an appropriately controlled error rate by data filtration, the use of our targeted hybrid library to process DIA data can substantially deepen the coverage for a given protein superfamily beyond conventional proteomic data acquisition and processing techniques.

\textbf{Validation of novel GPCR peptides exclusively identified with the hybrid library}

Compared to the max ID lists, there were 412 non-redundant novel GPCR peptides identified in three mouse brain regions after data filtration (Table S6). These peptides represent novel identifications beyond any DIA- or DDA-based experimental measurement and they were only obtained using our DIA hybrid library strategy. Next, we designed experiments to empirically validate their existence in the protein digests of the specific brain regions (Fig. 3A). We first conducted targeted MS assays on novel peptides in each region by parallel-reaction-monitoring (PRM) analyses. High-quality MSMS spectra were obtained for 214 peptides (Table S7). When
we incorporated these experimental fragment ion intensities and measured iRTs into the initial DIA spectral library, searching the DIA data with Spectronaut default parameters verified the identification of 207 novel peptides in the mouse brain tissues (Table S7). Thus, 50.2% of peptides identified exclusively using our DIA hybrid library approach are indeed present in the samples. Given that individual biological replicates were prepared for the DIA and PRM experiments, we speculate that the invalidated portion may result from relatively low reproducibility of extracting and analyzing these novel peptides most of which are at low abundance.

Still pursuing validation of the remaining 205 putative novel GPCR peptides, we searched for synthetic peptides of the same sequences deposited to the ProteomeTools spectral libraries (Zolg et al., 2017). The MSMS spectra retrieved for the 23 synthetic peptides were incorporated into the initial DIA spectral library for DIA data searching, which resulted in the validation of 22 additional novel GPCR peptides (Table S7). Thus, a total of 221 non-redundant novel peptides were validated with experimental MSMS data (Fig. 3B), corresponding to a total validation rate of 53.6% (Table S6). In detail, the GPCR peptide validation rates achieved with our DIA hybrid library strategy in the three brain regions are 55.5%, 44.8%, and 51.9%, respectively (Fig. 3C). These validated novel peptides confirmed the existence of 12 out of 22 novel GPCR proteins only discovered with the DIA hybrid library (Fig. 3B). Mirror plots for two example peptides comparing the predicted fragmentation patterns alongside experimental MSMS spectra from either the PRM analysis or the synthetic human proteome database reflect strong agreement between prediction and measurement. (Figs. 3D-3E).

**Deepening the proteome coverage for other transmembrane protein families or for GPCRs using a published dataset**

To extend our strategy to deepen coverage for transmembrane proteins of other families, we employed the same workflow described in Fig. 1 to re-train the deep learning models with
sample-specific DIA data and to construct hybrid spectral libraries (under the P2 optimal condition) for the 240 ion channels and the 452 transporters encoded by the mouse genome (Figure S9). These protein family-targeted virtual libraries (used in concert with the stringent data filtration) enabled a substantial increase of the proteome coverage for both ion channels (from 89 proteins and 1,179 peptides identified in the three regions with the initial DIA library to 123 proteins and 2,335 peptides with the hybrid library) and transporters (from 195 proteins and 2,350 peptides identified with the initial DIA library to 268 proteins and 3,716 peptides with the hybrid library) (Figs. 4B, 4C, 4E, 4F; Table S8).

To avoid re-training models using different datasets for different protein families, we built generic models of pDeep and DeepRT with 90% of randomly selected precursors from the initial DIA library to construct virtual libraries for three transmembrane protein families (Figure S10). The generic models showed very similar performance to the protein family-specific models. Furthermore, DIA hybrid libraries generated based on prediction by the generic models yielded peptide identifications for three transmembrane protein families very comparable to the family-specific models (Figure S10). Therefore, it is feasible to build generic models from an experimentally acquired DIA dataset to predict for any selected protein family.

Finally, seeking to demonstrate the effectiveness of our strategy for mining previously acquired DIA data from other sources, we downloaded a published DIA dataset that was acquired from total lysates of the brain barrel cortex regions from newborn mice (Bruderer et al., 2017). We used the same procedure for model re-training and processing of the DIA dataset obtained at 4 time points each in technical triplicate (12 analyses in total) (Figure S11). Retrained models were implemented to construct a targeted virtual library for all mouse GPCRs with the P2 condition, which was then incorporated with the initial spectral library generated from the DIA dataset to yield a hybrid library. Compared to 27 GPCR proteins and 229 peptides in total identified with DIA data alone over the time course, our DIA hybrid library strategy
enabled exceptionally deep mapping of 83 GPCRs based on spectral matches of 895 peptides after data filtration (Fig. 5, Table S9).

**Discussion**

Compared to previous studies that demonstrated the applicability of deep learning-based models to both DDA and DIA data analysis (Gessulat et al., 2019; Tiwary et al., 2019), our study explores a fundamentally new direction for DIA data mining. In previous work, a virtual spectral library was first generated for all experimentally detected peptides from a sample-specific spectral library derived from DDA analysis. Subsequent replacement of the entire experimental spectral library with the virtual library achieved a whole-proteome coverage close to the experimental library. In contrast, our study aims to deepen the sub-proteome coverage for a selected protein family that surpasses experimental limits. We constructed a hybrid spectral library that supplemented a DIA experiment-derived library with a protein family-targeted virtual library built on all *in silico* digested peptide precursors.

Our study demonstrates how DIA data mining using this hybrid library strategy can remarkably increase the depth of proteomic profiling for a selected protein family without compromising FDR control or quantification reproducibility. As a proof-of-concept, GPCR identifications in three mouse brain regions were increased from 310 peptides mapped to 38 proteins with the initial DIA library (using default software parameters) to 810 peptides mapped to 71 proteins with the DIA hybrid library (after stringent data filtration). Moreover, 412 novel GPCR peptides and 22 novel GPCR proteins were not observed in any conventional mining of data from DIA or pre-fractionation-based DDA experiments. Importantly, we performed orthogonal PRM experiments and exploited the synthetic human proteome database to validate the existence of 53.6% novel GPCR peptides detected exclusively with our DIA hybrid library strategy.
It is important to emphasize that our workflow for deep learning-based virtual library construction and subsequent generation of hybrid libraries only requires a small DIA dataset, which can be acquired from minimal sample quantities and minimal instrument time. Thus, our DIA hybrid library approach circumvents the need for the time-consuming generation of sample-specific spectral libraries through extensive pre-fractionation of peptide samples. Furthermore, this strategy is adaptable to targeting multiple protein families and to previously acquired DIA datasets, allowing for retrospective data mining based on new hypotheses. Although not directly showcased here, we anticipate that this bioinformatics strategy will enhance the detection of members in any protein family (or any set of selected proteins) of biological interest, including splicing isoforms, sequence variants, or proteins bearing specific PTMs, to an unprecedented depth.

**Limitations of the Study**

Our current study deepens the sub-proteome coverage for three transmembrane protein families based on the DIA-MS dataset acquired from mouse brain tissues. It remains to be investigated whether we can exploit this DIA hybrid library strategy to map other protein families from the same dataset in largely increased depth. Moreover, as we need to construct virtual libraries for individual pre-determined protein families, it would be more efficient to combine them and build a larger hybrid library for DIA-MS data search. But we also recognize the challenge of FDR control with an expanded hybrid library. Thus it is necessary to further develop the decoy library method for accurate assessment of FDR in DIA-MS data search with a hybrid spectral library.
**Data availability**

The MS data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the iProX partner repository (Ma et al., 2019) under the accession number PXD016441 and with the project ID IPX0001881000.

**Code availability**

The pipeline (TargetDIA) and relevant scripts for virtual library generation and DIA data search are available on GitHub at https://github.com/ComatoseHamster/TargetDIA.
Acknowledgements

We very much thank the fruitful discussion with Prof. Simin He and Dr. Wenfeng Zeng from Institute of Computing Technology, CAS, and their help with using pDeep 2. This work was funded by ShanghaiTech University, the National Program on Key Basic Research Project of China (2018YFA0507004 (W.S.), 2016YFA0501900 (Y.Z.), 2016YFA0501904 (Y.Z.), 2016YFC0905900 (S.Z.), 2018YFA0507000 (S.Z.)), and National Natural Science Foundation of China (31971362, 31671428, 31971178).

Author contributions

R.L. and K.D. performed model re-training, library generation and MS data processing. P.T. prepared samples, acquired and analyzed PRM data. S.L. prepared samples and acquired DIA data. C.T. and Y.L. helped with brain sample preparation and MS data acquisition. S.Z. and Y.Z. were involved in the overall project management. W.S., R.L. and P.T. wrote the manuscript with edits from all authors. W.S. conceived and supervised the project.

Declaration of Interests

The authors declare no competing financial interests.
Reference

Bekker-Jensen, D.B., Kelstrup, C.D., Batth, T.S., Larsen, S.C., Haldrup, C., Bramsen, J.B., Sorensen, K.D., Hoyer, S., Orntoft, T.F., Andersen, C.L., et al. (2017). An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. Cell Syst 4, 587-599 e584.

Bruderer, R., Bernhardt, O.M., Gandhi, T., Miladinovic, S.M., Cheng, L.Y., Messner, S., Ehrenberger, T., Zanotelli, V., Butscheid, Y., Escher, C., et al. (2015). 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.

Bruderer, R., Bernhardt, O.M., Gandhi, T., Xuan, Y., Sondermann, J., Schmidt, M., Gomez-Varela, D., and Reiter, L. (2017). Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. Mol Cell Proteomics 16, 2296-2309.

Gessulat, S., Schmidt, T., Zolg, D.P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., et al. (2019). Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods 16, 509-518.

Gillet, L.C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics 11, O111 016717.

Huang, Y., Todd, N., and Thathiah, A. (2017). The role of GPCRs in neurodegenerative diseases: avenues for therapeutic intervention. Curr Opin Pharmacol 32, 96-110.

Katritch, V., Cherezov, V., and Stevens, R.C. (2013). Structure-function of the G protein-coupled receptor superfamily. Annu Rev Pharmacol Toxicol 53, 531-556.

Kusebauch, U., Campbell, D.S., Deutsch, E.W., Chu, C.S., Spicer, D.A., Brusniak, M.Y., Slagel, J., Sun, Z., Stevens, J., Grimes, B., et al. (2016). Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome. Cell 166, 766-778.

Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B.C., and Aebersold, R. (2018). Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Mol Syst Biol 14, e8126.

Ma, C., Ren, Y., Yang, J., Ren, Z., Yang, H., and Liu, S. (2018). Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. Anal Chem 90, 10881-10888.

Ma, J., Chen, T., Wu, S., Yang, C., Bai, M., Shu, K., Li, K., Zhang, G., Jin, Z., He, F., et al. (2019). iProX: an integrated proteome resource. Nucleic Acids Res 47, D1211-D1217.

Pan, S.J., and Yang, Q. (2010). A Survey on Transfer Learning. IEEE Trans Knowl Data Eng 22, 1345 - 1359.

Picotti, P., and Aebersold, R. (2012). Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat Methods 9, 555-566.

Rosenberger, G., Bludau, I., Schmitt, U., Heusel, M., Hunter, C.L., Liu, Y., MacCoss, M.J., MacLean, B.X., Nesvizhskii, A.I., Pedrioli, P.G.A., et al. (2017). Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. Nat Methods 14, 921-927.

Schubert, O.T., Gillet, L.C., Collins, B.C., Navarro, P., Rosenberger, G., Wolski, W.E., Lam, H., Amodei, D., Mallick, P., MacLean, B., et al. (2015). Building high-quality assay libraries for targeted analysis of SWATH MS data. Nat Protoc 10, 426-441.

Suski, J.M., Lebiedzinska, M., Wojtala, A., Duszynski, J., Giorgi, C., Pinton, P., and Wieckowski, M.R. (2014). Isolation of plasma membrane-associated membranes from rat liver. Nat Protoc 9, 312-322.

Ting, Y.S., Egerton, J.D., Bollinger, J.G., Searle, B.C., Payne, S.H., Noble, W.S., and MacCoss, M.J. (2017). PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. Nat Methods 14, 903-908.
Ting, Y.S., Egertson, J.D., Payne, S.H., Kim, S., MacLean, B., Kall, L., Aebersold, R., Smith, R.D., Noble, W.S., and MacCoss, M.J. (2015). Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. Mol Cell Proteomics 14, 2301-2307.

Tiwary, S., Levy, R., Gutenbrunner, P., Salinas Soto, F., Palaniappan, K.K., Deming, L., Berndl, M., Brant, A., Cimermancic, P., and Cox, J. (2019). High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. Nat Methods 16, 519-525.

Tsou, C.C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A.C., and Nesvizhskii, A.I. (2015). DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat Methods 12, 258-264, 257 p following 264.

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al. (2014). Mass-spectrometry-based draft of the human proteome. Nature 509, 582-587.

Wisniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nat Methods 6, 359-362.

Zhou, X.X., Zeng, W.F., Chi, H., Luo, C., Liu, C., Zhan, J., He, S.M., and Zhang, Z. (2017). pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. Anal Chem 89, 12690-12697.

Zolg, D.P., Wilhelm, M., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., Bailey, D.J., Gessulat, S., Ehrlich, H.C., Weininger, M., et al. (2017). Building ProteomeTools based on a complete synthetic human proteome. Nat Methods 14, 259-262.
Figure Legends

Fig. 1 Overall workflow for DIA data mining using a hybrid spectral library. This hybrid library is constructed by merging a DIA experiment-derived library with a protein family-targeted virtual library built on all in silico digested peptide precursors. The targeted virtual library is generated using two deep learning models (pDeep and DeepRT) which are re-trained by transfer learning based on sample-specific DIA data alone. This study examines protein extracts of membrane fractions prepared from three mouse brain regions (cerebellum, midbrain, and spinal cord). ID, identification; pred, predicted; exp, experimental.

Fig. 2 Increasing the depth of GPCR identification by DIA-MS with the hybrid library. A, In silico digestion of the 524 mouse GPCR proteins with 12 combinations (P1-P12) of peptide charge states, numbers of missed cleavages, and Met oxidation status. The peptide length is restricted to 7-33 residues. B, The number of peptide precursors in the initial DIA spectral library, and each "hybrid library" comprising the initial DIA spectral library plus the GPCR-targeted virtual library. C-D, Number of GPCR peptide identifications (IDs) (C) and protein IDs (D) in the cerebellum between the initial DIA library and 12 hybrid libraries. Relative to the number of peptide/protein IDs obtained with the initial DIA library (shown on the left), the proportion of shared IDs for each hybrid library is shown in blue, gained IDs in orange, and lost IDs in grey. The protein/peptide ID number in each fraction is annotated. Note that additional DDA experiments were conducted on pre-fractionated peptide samples from each brain region, and the GPCR identification lists from the initial DIA and new DDA experiments were merged to generate a "max ID list". Max ID recovery rates refer to the percentages of bona fide identifications from these max ID lists, which were recovered using our 12 hybrid libraries. E-F, The number of GPCR peptide (E) or protein (F) IDs using the decoy hybrid library (left) or the P2 hybrid library (right) with default Spectronaut parameters (Default) or after data filtration based on Cscore >0.9 (Filtered). Relative to the protein/peptide IDs with the initial DIA library
(left), the proportion of shared IDs is shown in blue, decoy IDs in green, and gained IDs in orange.

**Fig. 3 Validation of GPCR peptides exclusively identified with the DIA hybrid library strategy.** A, Workflow of novel peptide validation, based on targeted MS (PRM) analysis or on searching the synthetic human proteome database in ProteomeTools. Novel GPCR peptides were validated through DIA data searching of experimental MSMS spectra acquired from the PRM experiment or the synthetic peptide spectral repository. B, Total number of novel GPCR proteins or peptides identified in the three brain regions, validated based on the PRM experiment (red) or the synthetic peptide spectral repository (pink). Also shown are the remaining unconfirmed peptides (beige). C, Total number of novel GPCR peptides identified in each brain region using our DIA hybrid library strategy (beige), as well as the number validated based on the PRM experiment (red) or the synthetic peptide spectral repository (pink). D, Pseudo mirror plots of example MSMS spectra comparing the predicted (upper) and experimental (lower) fragmentation patterns for two novel GPCR peptides. The experimental MSMS spectrum was acquired from the PRM experiment for peptide EAGGLCIAQSVR (left) or from the synthetic reference spectral repository for peptide SVYVDDDSEAAGNR (right). The Pearson correlation coefficient (PCC) and normalized spectral contrast angle (SA) are indicated. PRM, parallel reaction monitoring.

**Fig 4. Comparison of GPCR, ion channel and transporter identifications (IDs) in mouse brain tissues using the initial DIA library (with default software settings) and the hybrid library P2 (after data filtration).** A-C, Comparison of GPCR (A), ion channel (B), and transporter (C) peptide IDs (upper) and protein IDs (lower) in three brain regions. D-F, Comparison of total non-redundant GPCR (D), ion channel (E), and transporter (F) peptide IDs (upper) and protein IDs (lower) identified from three regions. In each panel, relative to the
protein/peptide IDs with the initial DIA library (shown on the left), the proportion of shared IDs for hybrid library P2 is shown in blue, and gained IDs in orange. The number of protein IDs in each fraction is annotated.

**Fig. 5** DIA hybrid library substantially increases GPCR-targeted proteome coverage in the analysis of a published dataset. **A,** Comparison of GPCR peptide (upper) and protein (lower) IDs in mouse brain cortex at four time points (P9, P15, P30, P54) during development. **B,** Comparison of the total non-redundant GPCR peptide (upper) and protein (lower) IDs identified in this time course. Relative to the number of protein/peptide IDs with the initial DIA library (with default software parameters; left), the proportion of shared IDs for the P2 hybrid library (after data filtration; right) is shown in blue and gained IDs in orange. The number of protein IDs in each fraction is annotated.
Content of Supplementary Tables

Table S1. DIA data search report with the initial DIA library (related to Figure 1)
Table S2. Hybrid library basic information (related to Figure 2)
Table S3. DIA data search reports with 12 different hybrid libraries (related to Figure 2)
Table S4. DDA data search report (related to Figure 2)
Table S5. DIA data search report with the GPCR decoy library (related to Figure 2)
Table S6. List of 412 novel GPCR peptides (related to Figure 3)
Table S7. Novel GPCR peptide validation results (related to Figure 3)
Table S8. DIA data search reports for the other two transmembrane protein families (related to Figure 4)
Table S9. DIA data search report for the barrel cortex dataset (related to Figure 5)
### Table: Number of Peptide IDs

| GPCR | Cerebellum | Midbrain | Spinal cord | Total |
|------|------------|----------|-------------|-------|
| DIA  | 304        | 306      | 293         | 310   |
| Hybrid | 290       | 286      | 260         | 306   |

| Ion channel | Cerebellum | Midbrain | Spinal cord | Total |
|-------------|------------|----------|-------------|-------|
| DIA         | 1,166      | 1,106    | 1,023       | 1,179 |
| Hybrid      | 1,130      | 966      | 869         | 1,177 |

| Transporter | Cerebellum | Midbrain | Spinal cord | Total |
|-------------|------------|----------|-------------|-------|
| DIA         | 2,340      | 2,332    | 2,295       | 2,350 |
| Hybrid      | 2,259      | 2,246    | 2,181       | 2,346 |

- Gain
- Shared
Highlights

A virtual library is built for a selected protein family using deep learning models.

The hybrid library strategy vastly deepens the coverage for the targeted protein family.

53.6% of novel GPCR peptides identified with the DIA hybrid library are validated.

Extend the strategy to deep mapping of multiple transmembrane protein families.