https://doi.org/10.1093/nargab/lqad022

Characterization of proteome profile data of chemicals based on data-independent acquisition MS with SWATH method

Hiromu Ishiguro[1,†], Tadahaya Mizuno[1,*,†], Yasuo Uchida[2,*,†], Risa Sato[2], Hayate Sasaki[2], Shumpei Nemoto[1], Tetsuya Terasaki[2] and Hiroyuki Kusuhara[1,*]

[1]Graduate School of Pharmaceutical Sciences, the University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan and [2]Graduate School of Pharmaceutical Sciences, Tohoku University, Aoba, Aramaki, Aoba-ku, Sendai 980-8578, Japan

Received October 19, 2022; Revised February 06, 2023; Editorial Decision February 14, 2023; Accepted February 20, 2023

ABSTRACT

Transcriptomic data of cultured cells treated with a chemical are widely recognized as useful numeric information that describes the effects of the chemical. This property is due to the high coverage and low arbitrariness of the transcriptomic data as profiles of chemicals. Considering the importance of posttranslational regulation, proteomic profiles could provide insights into the unrecognized aspects of the effects of chemicals. Therefore, this study aimed to address the question of how well the proteomic profiles obtained using data-independent acquisition (DIA) with the sequential window acquisition of all theoretical mass spectra, which can achieve comprehensive and arbitrariness-free protein quantification, can describe chemical effects. We demonstrated that the proteomic data obtained using DIA-MS exhibited favorable properties as profile data, such as being able to discriminate chemicals like the transcriptomic profiles. Furthermore, we revealed a new mode of action of a natural compound, harmine, through profile data analysis using the proteomic profile data. To our knowledge, this is the first study to investigate the properties of proteomic data obtained using DIA-MS as the profiles of chemicals. Our 54 (samples) × 2831 (proteins) data matrix would be an important source for further analyses to understand the effects of chemicals in a data-driven manner.

INTRODUCTION

Transcriptomic data of cultured cells treated with a specific chemical have been used as numeric information that describes the effects of the chemical. The aggregation of such profiles, hereafter referred to as profile data, has been employed as a database for searching similar chemicals, such as that in the connectivity map (CMap) project, or subjected to latent variable models to identify essential relationships between genes, which has contributed to the understanding of the effects of chemicals (1–4). Because it is necessary to capture the responses of cells comprehensively and describe the effects of chemicals without arbitrariness, the choice of using the transcriptome is rather reasonable. However, the importance of posttranslational regulation is evident in molecular biology, and there may be effects of chemicals that cannot be fully described using the transcriptomic layer.

As one of the other methods to obtain profiles of chemicals, a morphology-based approach has been developed by leveraging high content analyzers (5,6). While this approach is attractive because of its extremely high throughput nature, it is difficult to interpret individual features using this method. Although the approach works well for evaluating the similarity of chemicals, subsequent analyses are difficult to perform for understanding the mechanisms from the viewpoint of molecular biology. Similarly, the 2D electrophoresis approach has been used with success in identifying novel aspects of chemicals, but it is difficult to interpret the features biologically using this approach (7,8). In terms of interpretability, the proteomic approach is superior because, as with the transcriptome, the features are specific protein names and can be directly interpreted biologically. For instance, Creech *et al*. developed a proteomic profile platform called the Global Chromatin Profiling by measuring global modifications to histones and Abelin *et al.* developed a proteomic profile platform called the P100 by measuring approximately 100 phosphosites of proteins, both of which constitute the library of integrated network-based cellular signatures (LINCS) project (9,10). However, in both cases, coverage has been an issue because the number of

[*]To whom correspondence should be addressed. Tel: +81 3 5841 4771; Fax: +81 3 5841 4766; Email: tadahaya@mol.f.u-tokyo.ac.jp
Correspondence may also be to Yasuo Uchida. Tel: +81 22 795 6832; Fax: +81 22 795 6886; Email: yasuo.uchida.c8@tohoku.ac.jp
Correspondence may also be to Hiroyuki Kusuhara. Tel: +81 3 5841 4770; Fax: +81 3 5841 4766; Email: kusuhara@mol.f.u-tokyo.ac.jp

[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

variables is on the order of 10^2, which is less than that of successful transcriptome, and the descriptive ability of chemical effects depends on the properties of the scope of the focused proteome although these platforms capture irreplaceable layers for describing chemical effects. Therefore, there are few methodologies for quantifying compound responses that concurrently satisfy both comprehensiveness and interpretability.

Recent technical advances in proteome with MS enables simultaneous acquisition of 10^3 order or even more proteins such as data-dependent acquisition with Tandem Mass Tag technique and data-independent acquisition (DIA) with the sequential window acquisition of all theoretical mass spectra (SWATH-MS) method (11–14). The proteomic data acquired using these methods are not restricted to specific proteomic scopes and would have ideal properties as a chemical profile contributing to the understanding of the effects of chemicals. However, this hypothesis has yet to be tested. Therefore, the present study aimed to test the usefulness of the proteomic profiles obtained using DIA-MS without focusing on specific proteome scopes in understanding the chemical effects using profile data analysis. To our knowledge (based on a survey of around 300 studies in the PubMed searched using a query, ((data-independent acquisition) OR (DIA) OR (SWATH-MS)) AND (proteome) AND (drug OR chemical) in December 2022), this is the first study to address the question of how well the proteomic profiles obtained using DIA-MS can describe chemical effects, although the methodology is often employed for investigating the effects of a specific chemical (15). The overall design of the present study is shown in Figure 1. We found that the proteomic profile data were comparable with the transcriptomic profiles regarding the properties discriminating the effects of compounds and captured a novel aspect of a natural product, harmine, which could not be recognized in the analyses of transcriptomic profiles. The profile data set obtained in this study, although relatively small in size, has few confounding factors and could be an important source for understanding the properties of chemical profiles in the proteomic layer.

MATERIALS AND METHODS

Cell culture

MCF7 cells were cultured in Dulbecco modified Eagle's medium (11995–065, Life Technologies, Carlsbad, CA, USA) with 10% fetal bovine serum. All cells were maintained at 37°C under 5% CO_2.

Treatment of chemicals

The list of chemicals used in this study is presented in Supplementary Table S1. MCF7 cells were treated with the indicated chemical for 24 h at 37 °C. The treatment concentration was determined based on the morphological changes, as shown in Supplementary Figure S1.

Preparation of whole cell sample for proteomic analysis

Cells were seeded in six-well plates at a density of 1.0×10^5 cells/well and maintained for 48 h. After drug treatment, cell morphology was evaluated using a microscope. Then, the cells were washed twice using phosphate buffered saline and collected using a cell scraper. Pelleted cells were stored in an −80°C freezer until further use.

Protein digestion using lysyl endopeptidase and trypsin

Protein digestion was performed as described previously (16). Briefly, 40–50 μg protein obtained from MCF7 cells was solubilized in a denaturing buffer (7 M guanidium hydrochloride, 0.5 M Tris–HCl [pH 8.5], and 10 mM EDTA). The solubilized proteins were reduced using dithiothreitol for 1 h at 25 °C and subsequently S-carboxymethylated using iodoacetamide for 1 h at 25 °C. The alkylated proteins were precipitated using a methanol-chloroform-water mixture. The precipitates were solubilized using 6 M urea in 0.1 M Tris–HCl (pH 8.5) and diluted fivefold using 0.1 M Tris–HCl (pH 8.5) containing 0.05% ProteaseMax surfactant (Promega, Madison, WI, USA). The dilutions were incubated with lysyl endopeptidase (Lys-C; Wako Pure Chemical Industries, Osaka, Japan) at an enzyme:substrate ratio of 1:100 for 3 h at 30 °C. Subsequently, Lys-C digested proteins were treated with TPCK-treated trypsin (Promega) at an enzyme:substrate ratio of 1:100 for 16 h at 37 °C. The digested samples were cleaned up using a self-packed SDB-XD 200 μl tip (3M, Saint Paul, MN, USA).

Comprehensive quantitative protein expression profiling using SWATH-MS

Comprehensive quantitative protein expression profiles were obtained using SWATH-MS as described previously (17). Briefly, the cleaned peptide samples of MCF7 cells were injected into a NanoLC 425 system (Eksigent Technologies, Dublin, CA, USA) coupled with an electrospray-ionization Triple TOF 5600 mass spectrometer (SCIEX; Framingham, MA, USA), which was set up for a single direct injection, and analyzed using SWATH-MS acquisition. The peptides were directly loaded onto a self-packed C18 analytical column, which was prepared by packing ProntoSIL 200-3-C18 AQ beads (3 μm, 120 Å, Bischoff Chromatography, Germany) in a PicoFrit tip (ID 75 μm, PF360-75-10-N5, New Objective) with a length of 20 cm. After sample loading, the peptides were separated and eluted using a linear gradient; 98% A:2% B to 65% A:35% B (0–120 min), increased to 0% A:100% B (120–121 min), maintained at 0% A:100% B (121–125 min), reduced to 98% A:2% B (125–126 min), and then maintained at 98% A:2% B (126–155 min). Mobile phase A composition was 0.1% formic acid in water, and mobile phase B contained 0.1% formic acid in acetonitrile. The flow rate was 300 nl/min. The eluted peptides were positively ionized and measured in the SWATH mode. The measurement parameters were as follows: SWATH window, 64 variable windows from 400 to 1200 m/z; product ion scan range, 50–2000 m/z; declustering potential, 100; rolling collision energy value, $0.0625 \times [m/z \text{ of each SWATH window}] - 3.5$; collision energy spread, 15; and accumulation time, 0.05 s for each SWATH window. Spectral alignment and data extraction from the SWATH data were performed with the SWATH Processing Micro App in PeakView (SCIEX) using two

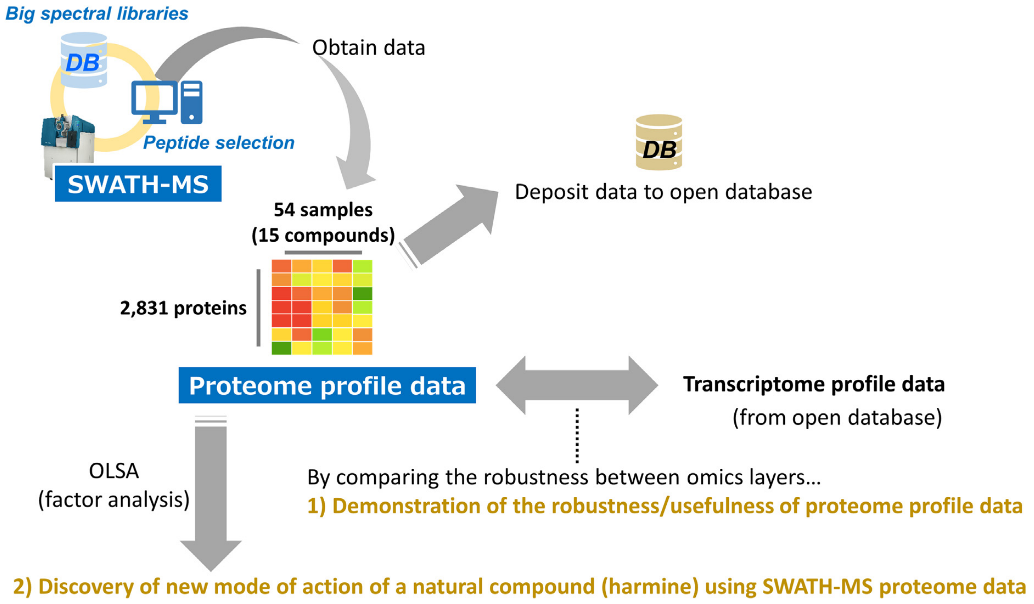

Big spectral libraries
DB
Obtain data
Peptide selection
SWATH-MS
DB
Deposit data to open database
54 samples
(15 compounds)
2,831 proteins
Proteome profile data
Transcriptome profile data
(from open database)
OLSA
(factor analysis)
By comparing the robustness between omics layers…
1) Demonstration of the robustness/usefulness of proteome profile data
2) Discovery of new mode of action of a natural compound (harmine) using SWATH-MS proteome data

Figure 1. Overall study design.

spectral libraries: an in-house spectral library and a publicly available pan-human library (18) for increasing the identification number of expressed proteins. The parameters for peak data extraction using PeakView were as follows: number of peptides per protein, 999; number of transitions per peptide, 6; peptide confidence threshold, 99%; false discovery rate threshold, 1.0%; extracted ion chromatogram extraction window, ±4.0 min; and extracted ion chromatogram width (ppm), 50. Unreliable data were excluded based on the in-house data/peptide selection workflow (19), and the quantification data of 2831 proteins in total were obtained in the final data set. It means that 2831 proteins were quantified in at least one sample. 1000 count of peak area was used for missing values in transition level.

Data processing

Each signal intensity of each fragment ion was standardized by dividing it with the mean value of all measured samples. After conversion to protein expression data, quantile normalization with the median of each rank data set was conducted to make all sample data identical regarding statistical properties. Cleaned data were converted to a modified robust z score as follows:

$$modified\,robust\,z = \frac{x - median}{1.4826 \times MAD}$$

$$MAD = median\,(|x - median|)$$

where x is each value of data, and median and median absolute deviation (MAD) are calculated using either only control sample data or all sample data. All data processing procedures were performed using the Pandas, NumPy, and SciPy packages in Python 3.7.

Transcriptomic profile data

The transcriptomic profile data were downloaded from the website of iLINCS (http://www.ilincs.org/ilincs/datasets/LINCS). We employed the profile data of MCF7 cells treated with the indicated chemicals. As for compounds that have multiple data with different treatment concentrations, averaged data were employed.

Factor analysis of profile data with OLSA

To extract latent variables in profile data, both processed proteomic and transcriptomic profiles are subjected to OLSA algorithm, which is a factor analysis adapted for chemical profiles considering reversibility of effects of chemicals such as agonism and antagonism and is available at https://github.com/mizuno-group/OLSApy (1–4). The number of factors is a hyperparameter and was determined by parallel analysis in this study (20). The number is 16 for the proteomic profiles and 9 for the transcriptomic profiles.

Gene ontology analysis

Gene ontology analysis was conducted using the data set derived from the Gene Ontology consortium (biological process, 2019). Fisher's exact test was employed for the calculation of enrichment, and the obtained P-values were adjusted using the Benjamini–Hochberg method.

Immunofluorescence and acquisition of signal intensity data

After the MCF7 cells were treated with each compound for 24 h, they were rinsed with PBS two times. Then, they were fixed by incubating with 4% paraformaldehyde for 10 min and made permeable by incubating with Triton-X for 5 min. Then, the cells were stained using an anti-coilin antibody (#ab11822, Abcam, Cambridge, UK) and visualized by confocal microscopy using a laser-scanning confocal microscope (TCS SP5 II; Leica, Solms, Germany) and a high content analyzer (ArrayScan VTI; Thermo Scientific, Waltham, MA, USA). Signal intensities were measured using the imaging software module of ArrayScan called BioApplications.

RESULTS

Selection of test compounds, acquisition of proteomic data and conversion to response profiles

First, we selected the compounds to be used in this study. To explore the usefulness of proteomic data as proteomic profiles, we decided to compare the characteristics of proteomic data with those of CMap transcriptomic data, which have already been confirmed to be useful in profile data analysis (21,22). Based on the preliminary analyses, we selected 15 compounds from eight groups that showed distinct separation in the result of orthogonal linear separation analysis (OLSA), which is a modified factor analysis we previously developed to delineate the multiple effects of chemicals (Supplementary Figure S2) (3). Proteomic data of the selected 15 compounds were obtained using DIA-MS with the SWATH-MS method, which resulted in a 54 (samples) × 2831 (proteins) data matrix. The expression data were converted into population-based modified robust *z*-score data (2) and averaged over compounds. Data quality was confirmed by visualizing heatmap of each sample (Supplementary Figure S3).

Evaluation of robustness against the random noise of proteomic profile data

Robustness against noise is one of the advantages of profile data analysis, which captures the covariance structure of data. To verify this advantage, 100 noise-added data sets were generated by adding random noise following the normal distribution with 0 as the average; 20%, 40%, 60%, 80%, 100%, 200% and 300% of the standard deviation (SD) calculated from proteomic profiles. Then, we investigated how much the correlation coefficient between each chemical profile and the other profiles changed after the Gaussian noise was added. Consequently, no clear difference was found between the proteomic and transcriptomic data (Figure 2A and B). Moreover, even when Gaussian noise equal to threefold of the SD of each sample data was added, the smallest correlation coefficients were 0.904 and 0.879 in the proteomic and transcriptomic profile data, respectively. This finding suggested that the acquired proteomic data were as robust against the noise as the transcriptomic data. Next, we investigated how much noise addition affects the factor structures and whether the degree of influence differs between the transcriptomic and proteomic data using noise-added data sets. After adding random noise, the correlation coefficients between the factors derived from profile data with or without random noise. Compared with whole variable analysis, the factor structure was largely affected by noise addition, and the smallest correlation coefficients in both data types were approximately 0.2 with 3 SD noise (Figure 2C and D). Regarding the layer difference, no clear difference was observed between the proteomic and transcriptomic data.

Evaluation of robustness against missing variables of proteomic profile data

In addition to robustness against noise, robustness against missing variables is also one of the advantages of profile data analysis. To verify this advantage, 100 data sets were generated by randomly deleting 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90% of the variables of both types of chemical profiles. Then, we investigated how much the correlation coefficient between each chemical profile and the other profiles changed when the variables were deleted. We found that even when 90% of the variables were deleted, the correlation coefficients were high (0.914 and 0.855 for the proteomic and transcriptomic profile data, respectively). This finding suggested that the proteomic data were as robust against missing variables as the transcriptomic data (Figure 3A and B). Then, we investigated how missing variables affect the factor structure and whether the degree of influence differs between the transcriptomic and proteomic data using the above data sets whose variables were randomly deleted. After deleting variables, the correlation coefficients between the factors derived from profile data with or without random noise. Consequently, the smallest correlation coefficient in both types of data was greater than 0.6, suggesting that the factor structures derived from both profile data were almost unchanged by missing variables (Figure 3C and D).

Profile data analysis in the proteomic layer

The findings suggesting the usefulness of proteomic profiles obtained using the proteomic profile data as profile data motivated us to elucidate novel aspects of chemicals by analyzing the proteomic profile data. Thus, we analyzed the proteomic profile data using OLSA and obtained 13 factors in this time. Notably, the proteomic profiles were subjected to the analysis without taking the average, unlike that in Figures 1–3, to capture the variances of the factor scores of each compound. To investigate whether the obtained factors have biological meanings, we performed a GO analysis for each set of the main constituent variables of the factors. We found that all 13 factors had one or more annotations at the significance level of 0.05, even after the *P*-values were adjusted using a multiple-test correction (Table 1). This was accomplished owing to the comprehensiveness of the proteomic data and has never been achieved with other proteomic data acquisition methods with limited comprehensiveness. Moreover, many chemicals had their characteristic factors associated with their known effects, as shown in the heat map of the response score matrix (Supplementary Figure S4). For example, trichostatin and vorinostat, which are both known as histone deacetylase inhibitors (23,24), were ranked high in the third factor (P3V) (Figure 4A), and cyclosporine and thapsigargin, which are both known to induce endoplasmic reticulum stress (ref), were ranked high in P5V, followed by geldanamycin, which is also known to induce endoplasmic reticulum stress (25) (Figure 4B).

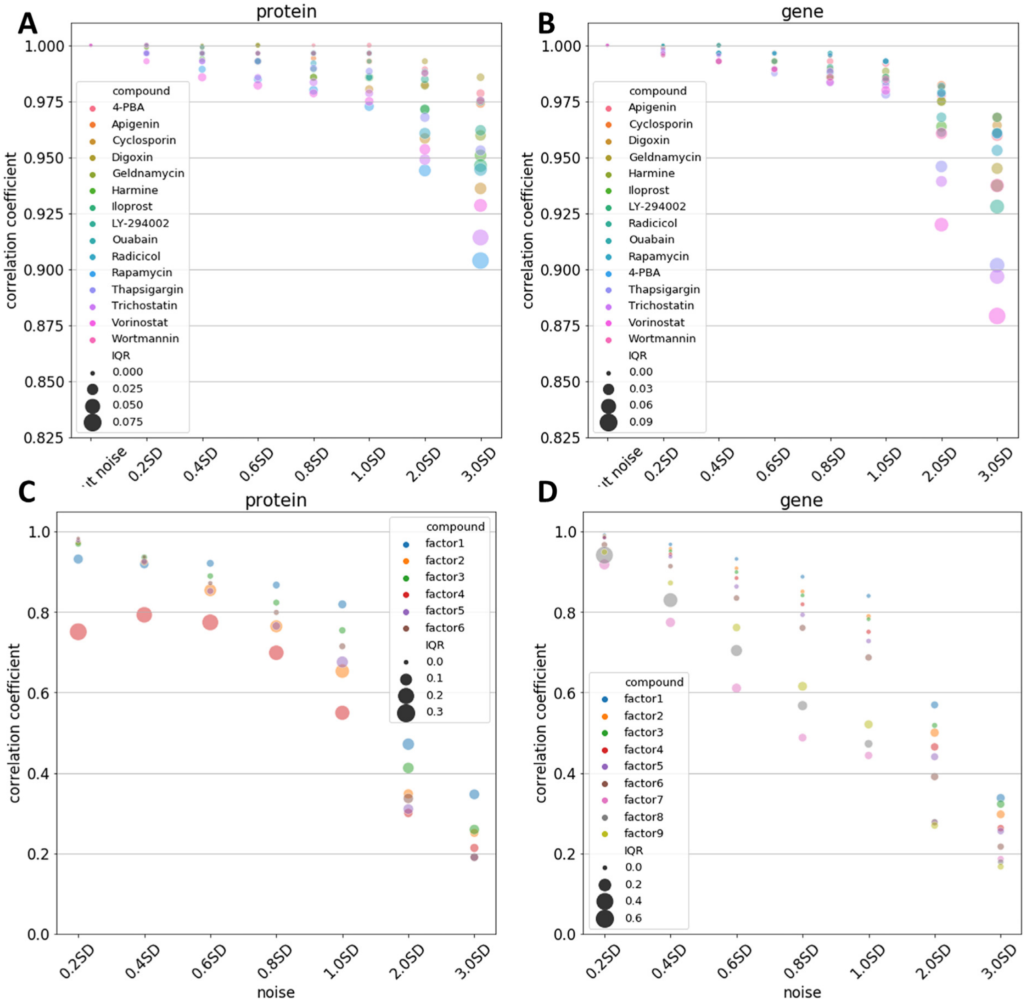


Figure 2. Robustness of the proteomic profile data against noise. (**A**, **B**) Bubble plots of correlation coefficients calculated using the SWATH-MS proteomic and CMap transcriptomic data, respectively. Each bubble indicates a median value and the IQR value of 100 rank-based correlation coefficients between the original rank order of each compound as the reference data and the rank order after Gaussian noise addition. (**C**, **D**) Bubble plots of correlation coefficients calculated using the SWATH-MS proteomic and CMap transcriptomic data, respectively. Each bubble indicates a median value and the IQR value of correlation coefficients between each factor data as the reference data and other factors. SD: standard deviation, IQR: interquartile range.

Discovery of a new mode of action of harmine based on the proteomic profile data

In the result of proteomic data analysis using OLSA, the ninth factor, P9V, was specific for harmine, which is a natural product (Supplementary Figure S5). In general, natural products have many effects and are often employed as seed compounds in drug discovery (22,26). Therefore, we looked into the detail of P9V. We found that the GO terms for P9V were related to telomeres or the Cajal body, which was not found in the analyses of the transcriptomic profiles (Supplementary Tables S2 and S3). The Cajal body, which was first discovered by Ramon y Cajal, is one of the spherical organelles found in the nuclei of proliferative cells (27). Although coilin is known as the marker protein of the Cajal body, no study has investigated the effect of harmine on coilin. Therefore, we investigated the effect of harmine on coilin using imaging analysis. An immunofluorescence study showed that harmine treatment increased the signals of coilin in the nuclei of MCF7 cells (Figure 5A). Thus, we quantitatively investigated the effects of harmine and apigenin, which was the second-highest scoring compound for P9V, on the expression levels of coilin using a high content analyzer. We found that harmine increased the expression levels of coilin, and that the degree of this increase was greater than that after apigenin treatment, which agreed with the P9V score (Figure 5B and Supplementary Figure S6).

DISCUSSION

As the term polypharmacology implies, it is widely accepted that a chemical has multiple effects (28). Latent variable models, such as factor analysis, of a data set composed of omics data of the cells treated with a variety of chemicals are an effective approach for understanding such multiple effects of a chemical, including even unrecognized ones (4). To maximize the power of the above combination, the utilized omics analysis method must be comprehensive and interpretable. Although transcriptomic data are the first

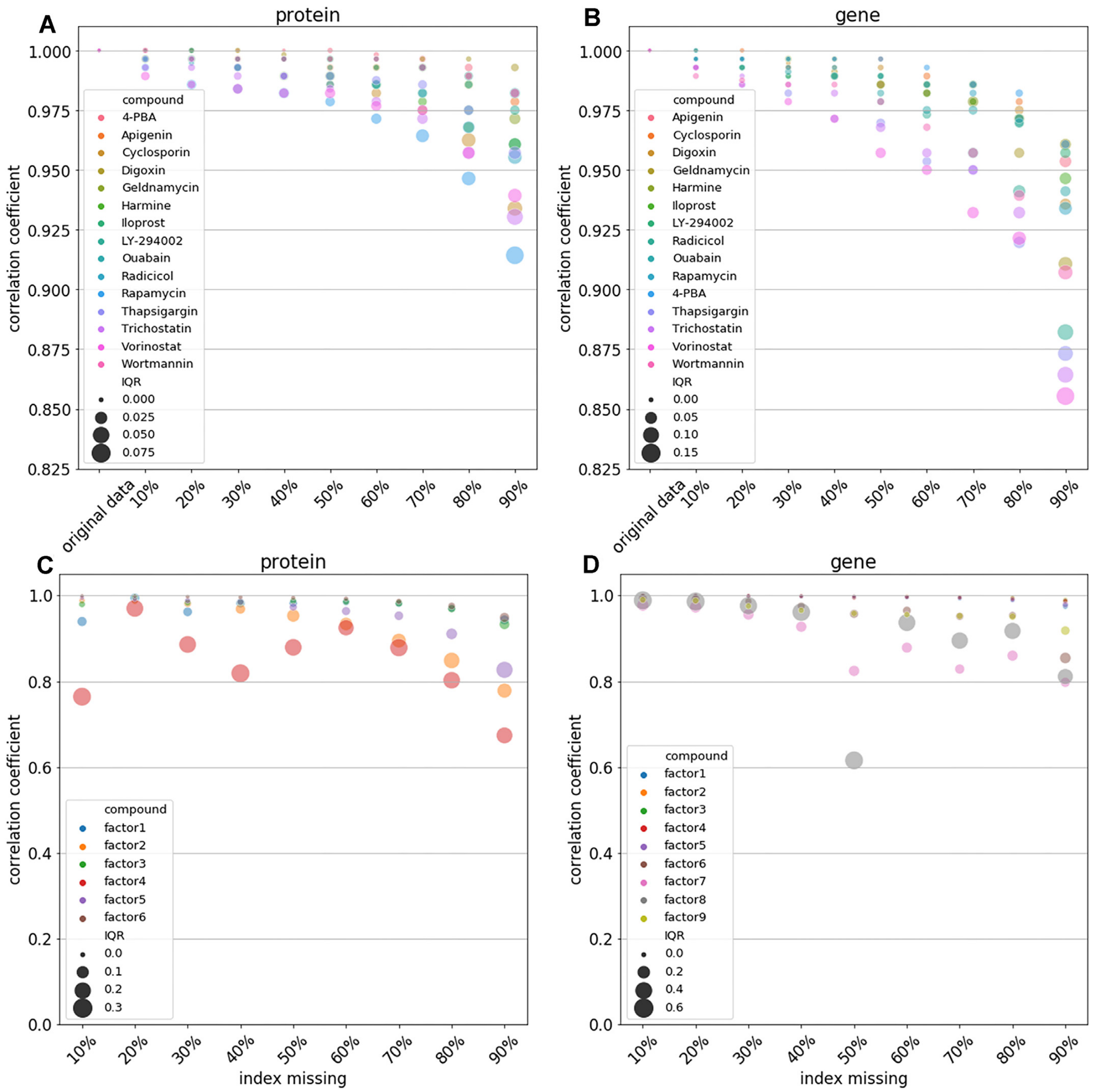


Figure 3. Robustness of the proteomic profile data against missing variables. (**A**, **B**) Bubble plots of correlation coefficients calculated using the SWATH-MS proteomic and CMap transcriptomic data, respectively. Each bubble indicates a median and the IQR value of 100 rank-based correlation coefficients between the original rank order of each compound as the reference data and the rank order after randomly deleting variables. (**C**, **D**) Bubble plots of correlation coefficients calculated using the SWATH-MS proteomic and CMap transcriptomic data, respectively. Each bubble indicates a median value and the IQR value of correlation coefficients between each factor data as the reference data and other factors. SD: standard deviation, IQR: interquartile range.

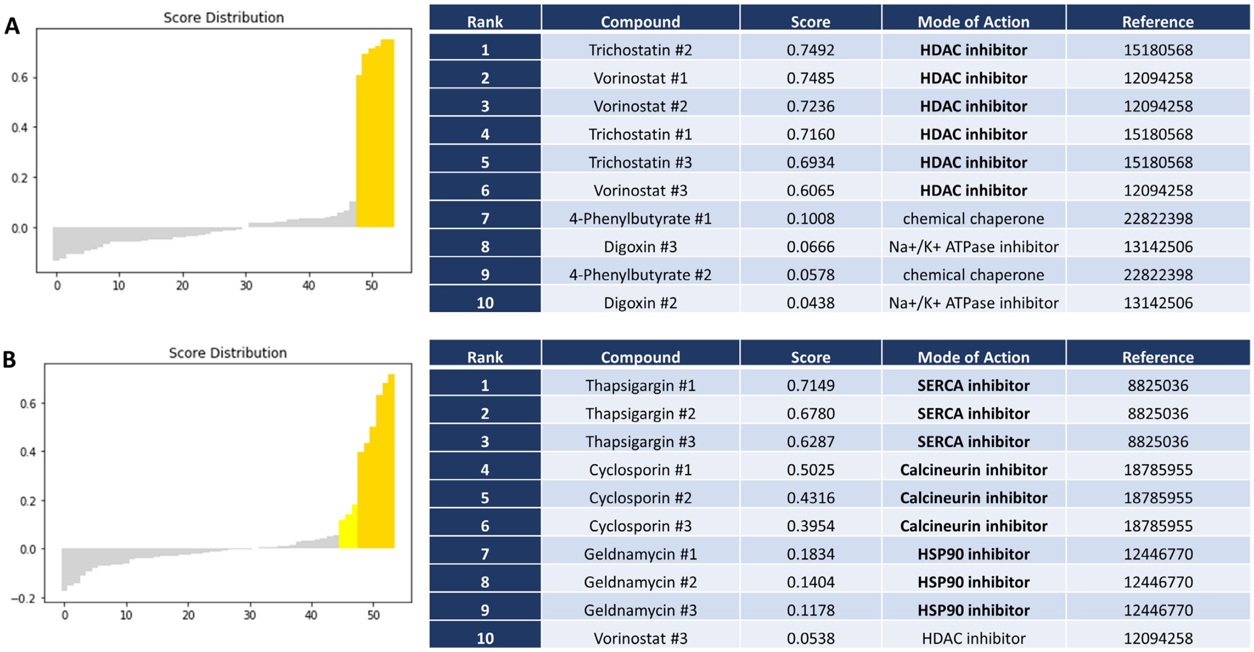
choice for the input of such profile data analysis, the characteristics of information obtained using omics data are different between different omics layers. In the present study, we investigated the usefulness of proteomic data obtained using DIA-MS, which satisfies the above two criteria, as the input for profile data analysis by: (i) acquiring and analyzing proteomic data obtained using the DIA-MS method, and (ii) comparing the obtained proteomic profiles with those in the CMap, which has been established as a valuable transcriptomic profile database (1). Based on a survey of 289 studies in PubMed searched using a query, ((data-independent acquisition) OR (DIA) OR (SWATH-MS)) AND (proteome) AND (drug OR chemical) in December 2022, this is the first study to address the question of how well the proteomic profiles obtained using DIA-MS can describe chemical effects.

First, we examined whether the proteomic data obtained using the DIA-MS method were robust against random noise and missing variables, considering the unstable nature of omics data. Note that we did not directly and quantitatively compare the transcriptome and proteome profiles of each drug in this study because it is impossible to reproduce specimens from other studies even if the same cell line and the same drug are used. We employed transcriptome profiles as a reference from public database and compared the two layers at the set level, focusing on a set of chemicals

Table 1. GO analysis of factors derived from proteomic profiles

Factor	GO term	*P*-value	Adjusted *P*-value	Hit proteins
Factor 1	Canonical glycolysis (GO: 0061621)	7.88E-13	1.23E-10	{‘gpi’, ‘pfkl’, ‘eno1’, ‘pkm’, ‘gapdh’, ‘pgam1’}
Factor 2	Branched-chain amino acid catabolic process (GO: 0009083)	1.98E-16	8.87E-14	{‘acat1’, ‘aldh6a1’, ‘mccc2’, ‘mccc1’, ‘hibadh’, ‘ivd’, ‘hmgcl’}
Factor 3	Fatty-acyl-CoA biosynthetic process (GO: 0046949)	1.21E-09	1.08E-06	{‘acsl4’, ‘scd’, ‘acly’, ‘acsl3’, ‘hsd17b12’}
Factor 4	Pre-mRNA cleavage required for polyadenylation (GO: 0098789)	2.51E-05	0.006586	{‘ncbp1’, ‘cpsf6’}
Factor 5	IRE1-mediated unfolded protein response (GO: 0036498)	2.02E-25	1.58E-22	{‘ssr1’, ‘srprb’, ‘gfpt1’, ‘hyou1’, ‘sec61b’, ‘pdia6’, ‘sec31a’, ‘mydgf’, ‘preb’, ‘dnajb11’, ‘hspa5’, ‘pdia5’, ‘sec63’, ‘arfgap1’, ‘srpra’}
Factor 6	Response to unfolded protein (GO: 0006986)	1.99E-22	1.70E-19	{‘hsph1’, ‘dnajb1’, ‘hspb1’, ‘hspa9’, ‘hsp90aa1’, ‘hspd1’, ‘serpinh1’, ‘hspa4l’, ‘hspa8’, ‘hsp90ab1’, ‘hspe1’, ‘dnaja1’}
Factor 7	Regulation of carbohydrate catabolic process (GO: 0043470)	5.11E-05	0.013868	{‘pgam1’, ‘nup43’, ‘nup88’}
Factor 8	SRP-dependent cotranslational protein targeting to membrane (GO: 0006614)	9.85E-09	2.30E-06	{‘rplp2’, ‘rpl18’, ‘rps25’, ‘rps19’, ‘rps7’, ‘rpl32’, ‘rps26’, ‘rplp1’}
Factor 9	Regulation of establishment of protein localization to telomere (GO: 0070203)	1.63E-08	7.63E-06	{‘cct2’, ‘cct7’, ‘cct4’}
Factor 10	Positive regulation of viral process (GO: 0048524)	3.32E-06	0.002116	{‘hacd3’, ‘ppib’, ‘polr2h’, ‘dhx9’}
Factor 11	Chylomicron assembly (GO: 0034378)	1.19E-05	0.004887	{‘apoe’, ‘apob’}
Factor 12	Actin filament capping (GO: 0051693)	1.19E-05	0.004578	{‘capzb’, ‘scin’}
Factor 13	Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay (GO: 0000184)	1.55E-13	8.82E-11	{‘rpl27a’, ‘rps6’, ‘rps11’, ‘rpl8’, ‘rpl31’, ‘smg1’, ‘rpl12’, ‘rpl27’, ‘rpl24’, ‘rpl34’, ‘rnps1’, ‘rpl21’, ‘rps20’}

Results of the GO analysis for main component variables of each factor. The top GO term of each factor is listed.



Rank	Compound	Score	Mode of Action	Reference
1	Trichostatin #2	0.7492	**HDAC inhibitor**	15180568
2	Vorinostat #1	0.7485	**HDAC inhibitor**	12094258
3	Vorinostat #2	0.7236	**HDAC inhibitor**	12094258
4	Trichostatin #1	0.7160	**HDAC inhibitor**	15180568
5	Trichostatin #3	0.6934	**HDAC inhibitor**	15180568
6	Vorinostat #3	0.6065	**HDAC inhibitor**	12094258
7	4-Phenylbutyrate #1	0.1008	chemical chaperone	22822398
8	Digoxin #3	0.0666	Na+/K+ ATPase inhibitor	13142506
9	4-Phenylbutyrate #2	0.0578	chemical chaperone	22822398
10	Digoxin #2	0.0438	Na+/K+ ATPase inhibitor	13142506

Rank	Compound	Score	Mode of Action	Reference
1	Thapsigargin #1	0.7149	**SERCA inhibitor**	8825036
2	Thapsigargin #2	0.6780	**SERCA inhibitor**	8825036
3	Thapsigargin #3	0.6287	**SERCA inhibitor**	8825036
4	Cyclosporin #1	0.5025	**Calcineurin inhibitor**	18785955
5	Cyclosporin #2	0.4316	**Calcineurin inhibitor**	18785955
6	Cyclosporin #3	0.3954	**Calcineurin inhibitor**	18785955
7	Geldnamycin #1	0.1834	**HSP90 inhibitor**	12446770
8	Geldnamycin #2	0.1404	**HSP90 inhibitor**	12446770
9	Geldnamycin #3	0.1178	**HSP90 inhibitor**	12446770
10	Vorinostat #3	0.0538	HDAC inhibitor	12094258

Figure 4. Profile data analysis in the proteomic layer. (**A**) Top 10 compounds for P3V factor and score distribution. P3V factor (the factor with the third highest contribution) scores of all compounds are arranged in ascending order and plotted on the graph. The rank, name, score, mode of action, and reference (PMID) of the top 10 compounds are shown. The top six compounds (trichostatin and vorinostat, which are well-known histone deacetylase inhibitors) are shown in gold. (**B**) Top 10 compounds for P5V factor and score distribution. P5V factor scores of all compounds are arranged in ascending order and plotted on the graph, as in (A). The top six compounds (thapsigargin and cyclosporin, which are well-known endoplasmic reticulum stress inducers) are shown in gold, and the next three compounds (geldanamycin) are shown in yellow.

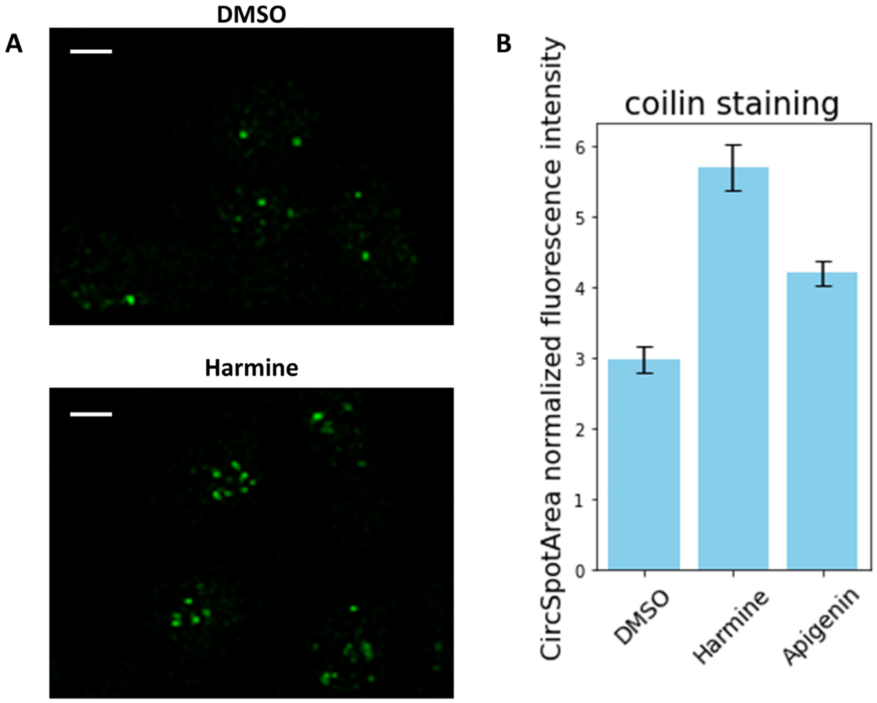


Figure 5. Discovery of a new mode of action of harmine based on the proteomic profile data. (**A**) Immunofluorescence analysis of coilin treated with or without harmine. Green signals indicate coilin in MCF7 cells treated with harmine at 16 μM for 24 h. Scale bars indicate 10 μm. (**B**) Increase of coilin signals in MCF7 cells by harmine or apigenin treatment. MCF7 cells were treated with harmine (16 μM) or apigenin (14.8 μM) for 24 h and stained using an anti-coilin antibody, and fluorescence signals were detected using Cellomics ArrayScan VTI. Black lines indicate the 95% confidence intervals.

with various effects. There were no clear differences in the responses to the addition of noise and removal of variables between the transcriptomic and proteomic profiles (Figures 2 and 3). Notably, the comparable performance of the proteomic profile data obtained using DIA-MS in the factor structure analysis indicates that these data capture the covariance structure caused by the analyzed chemicals as well as the transcriptomic profile data. Although the correlation coefficients cannot be directly compared because of the difference in the number of features, the fact that there was no clear difference in the response to a gradual increase in the perturbation intensity indicates that the proteomic profile data have a comparable ability to classify the chemicals as the transcriptomic profile data.

Our next concern was whether the analysis of the proteomic profile data contributed to the discovery of novel aspects of chemicals as the transcriptomic profile data. Although many reports of profile data analysis have discovered new effects of chemicals, no study has reported on proteomic profile data obtained using DIA-MS with the SWATH method. As shown in Figure 5, we found a novel feature of harmine, which was its relationship with the Cajal body, by focusing on the proteins that composed P9V. This could not be achieved without the high comprehensiveness and interpretability of DIA-MS-derived proteomic profiles. The Cajal body plays an important role in the biogenesis of small nuclear ribonucleoproteins, small Cajal body-specific ribonucleoproteins, small nucleolar ribonucleoproteins, and the telomerase, which are all crucial for efficient and rapid cell proliferation (27,29,30). One of the possible mechanisms underlying the association between harmine and the Cajal body may be dual-specificity tyrosine phosphorylation-regulated kinase 1A-dependent regulation because harmine is a potent inhibitor of this kinase (31). However, it should be noted that we cannot infer the causal relationship between them because the proteomic profile data is a snapshot of one time point. Although the relationship between dual-specificity tyrosine phosphorylation-regulated kinase 1A and Cajal bodies is not discussed in detail here, it is expected that further molecular biological experiments will reveal the molecular mechanisms by which harmine affects the Cajal body, which is an interaction recognized by the analysis of the proteomic profile data obtained using DIA-MS.

Although we selected compounds with as diverse effects as possible at the transcriptomic layer, one of the major limitations of the present study is that the number of compounds was limited. Recent methodological advances in the field such as Scanning SWATH greatly improve the throughput of data acquisition, in parallel with depth of protein coverage (32,33). Combination of these advanced methods with automated sample preparation enables acquisition of hundreds of proteomic chemical profiles and could clarify the differences between the transcriptomic and proteomic layers in larger scale than that of this study regarding describing the effects of chemicals.

CONCLUSION

Proteomic profile data obtained using DIA-MS with the SWATH method are as robust as transcriptomic data regarding describing a set of chemicals with various effects.

Owing to high comprehensiveness and interpretability, the proteomic profile data have the potential to enable the identification of novel aspects of chemicals, which would differ from those obtained using the transcriptomic layer.

DATA AVAILABILITY

The transcriptomic profile data used in this study are available in the iLINCS database (http://www.ilincs.org/ilincs/). The proteomic profile data (in both raw and processed forms), source codes, and codes for running analyses are packed in Supplementary Data.

The mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium via the PRIDE (34) partner repository with the data set identifier PXD032785.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank the investigators who obtained the transcriptomic data in the CMap project and curated them in the iLINCS database.

Author contributions: T.M. and Y.U. coordinated the project, conceived the concept, and designed the experiments. H.I., T.M., Y.U. and R.S. analyzed the data. H.I., R.S., H.S. and S.N. performed the experiments. H.I., T.M. and Y.U. wrote the manuscript. H.K. and T.T. revised the manuscript for intellectual content. All authors approved the manuscript before submission.

FUNDING

Grant-in-Aid for Challenging Exploratory Research [17K19478] from the Japan Society for the Promotion of Science (to H.K.); Japanese Society for the Promotion of Science for Scientific Research (B) [KAKENHI: 20H03399 to Y.U.]; Long-Range Research Initiative [17_S05-01-2] from the Japan Chemical Industry Association (H.K.); Grant-in-Aid from the Mochida Memorial Foundation for Medical and Pharmaceutical Research (to T.M.).

Conflict of interest statement. None declared.

REFERENCES

1. Lamb,J., Crawford,E.D., Peck,D., Modell,J.W., Blat,I.C., Wrobel,M.J., Lerner,J., Brunet,J., Subramanian,A., Ross,K.N. *et al.* (2006) The connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.
2. Subramanian,A., Narayan,R., Corsello,S.M., Peck,D.D., Natoli,T.E., Lu,X., Gould,J., Davis,J.F., Tubelli,A.A., Asiedu,J.K. *et al.* (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437–1452.
3. Mizuno,T., Kinoshita,S., Ito,T., Maedera,S. and Kusuhara,H. (2019) Development of orthogonal linear separation analysis (OLSA) to decompose drug effects into basic components. *Sci. Rep.*, **9**, 1824.
4. Mizuno,T., Morita,K. and Kusuhara,H. (2020) Interesting properties of profile data analysis in the understanding and utilization of the effects of drugs. *Biol. Pharm. Bull.*, **43**, 1435–1442.
5. Young,D.W., Bender,A., Hoyt,J., McWhinnie,E., Chirn,G.-W., Tao,C.Y., Tallarico,J.A., Labow,M., Jenkins,J.L., Mitchison,T.J. *et al.* (2008) Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.*, **4**, 59–68.
6. Bray,M.A., Singh,S., Han,H., Davis,C.T., Borgeson,B., Hartland,C., Kost-Alimova,M., Gustafsdottir,S.M., Gibson,C.C. and Carpenter,A.E. (2016) Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.*, **9**, 1757–1774.
7. Muroi,M., Kazami,S., Noda,K., Kondo,H., Takayama,H., Kawatani,M., Usui,T. and Osada,H. (2010) Application of proteomic profiling based on 2d-DIGE for classification of compounds according to the mechanism of action. *Chem. Biol.*, **17**, 460–470.
8. Kinoshita,S., Mizuno,T., Hori,M., Kohno,M. and Kusuhara,H. (2019) Development of a novel platform of proteome profiling based on an easy-to-handle and informative 2D-DIGE system. *Biol. Pharm. Bull.*, **42**, 2069–2075.
9. Abelin,J.G., Patel,J., Lu,X., Feeney,C.M., Fagbami,L., Creech,A.L., Hu,R., Lam,D., Davison,D., Pino,L. *et al.* (2016) Reduced-representation phosphosignatures measured by quantitative targeted MS capture cellular states and enable large-scale comparison of drug-induced phenotypes. *Mol. Cell. Proteomics*, **15**, 1622–1641.
10. Creech,A.L., Taylor,J.E., Maier,V.K., Wu,X., Feeney,C.M., Udeshi,N.D., Peach,S.E., Boehm,J.S., Lee,J.T., Carr,S.A. *et al.* (2015) Building the connectivity Map of epigenetics: chromatin profiling by quantitative targeted mass spectrometry. *Methods*, **72**, 57–64.
11. Li,K.W., Gonzalez-Lozano,M.A., Koopmans,F. and Smit,A.B. (2020) Recent developments in data independent acquisition (DIA) mass spectrometry: application of quantitative analysis of the brain proteome. *Front. Mol. Neurosci.*, **13**, 564446.
12. Gillet,L.C., Navarro,P., Tate,S., Röst,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.
13. Liu,D., Yang,S., Kavdia,K., Sifford,J.M., Wu,Z., Xie,B., Wang,Z., Pagala,V.R., Wang,H., Yu,K. *et al.* (2021) Deep profiling of microgram-scale proteome by tandem mass tag mass spectrometry. *J. Proteome Res.*, **20**, 337–345.
14. Aggarwal,S., Talukdar,N.C. and Yadav,A.K. (2019) Advances in higher order multiplexing techniques in proteomics. *J. Proteome Res.*, **18**, 2360–2369.
15. Nury,C., Merg,C., Eb-Levadoux,Y., Bovard,D., Porchet,M., Maranzano,F., Loncarevic,I., Tavalaei,S., Lize,E., Demenescu,R.L. *et al.* (2022) Toxicoproteomics reveals an effect of clozapine on autophagy in human liver spheroids. *Toxicol. Mech. Methods*, https://doi.org/10.1080/15376516.2022.2156005.
16. Uchida,Y., Zhang,Z., Tachikawa,M. and Terasaki,T. (2015) Quantitative targeted absolute proteomics of rat blood-cerebrospinal fluid barrier transporters: comparison with a human specimen. *J. Neurochem.*, **134**, 1104–1115.
17. Hellinen,L., Sato,K., Reinisalo,M., Kidron,H., Rilla,K., Tachikawa,M., Uchida,Y., Terasaki,T. and Urtti,A. (2019) Quantitative protein expression in the Human retinal pigment epithelium: comparison between apical and basolateral plasma membranes with emphasis on transporters. *Investig. Opthalmol. Vis. Sci.*, **60**, 5022.
18. Rosenberger,G., Koh,C.C., Guo,T., Röst,H.L., Kouvonen,P., Collins,B.C., Heusel,M., Liu,Y., Caron,E., Vichalkovski,A. *et al.* (2014) A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data*, **1**, 140031.
19. Uchida,Y., Sasaki,H. and Terasaki,T. (2020) Establishment and validation of highly accurate formalin-fixed paraffin-embedded quantitative proteomics by heat-compatible pressure cycling technology using phase-transfer surfactant and SWATH-MS. *Sci. Rep.*, **10**, 11271.
20. Ömay Çokluk,D.K. (2016) Using Horn's parallel analysis method in exploratory factor analysis for determining the number of factors. *Educ. Sci. Theory Pract.*, **16**, 537–551.
21. Morita,K., Mizuno,T. and Kusuhara,H. (2020) Decomposition profile data analysis of multiple drug effects identifies endoplasmic reticulum stress-inducing ability as an unrecognized factor. *Sci. Rep.*, **10**, 13139.

22. Nemoto,S., Morita,K., Mizuno,T. and Kusuhara,H. (2021) Decomposition profile data analysis for deep understanding of multiple effects of natural products. *J. Nat. Prod.*, **84**, 1283–1293.
23. Yoshida,M., Kijima,M., Akita,M. and Beppu,T. (1990) Potent and specific inhibition of mammalian histone deacetylase both in vivo and in vitro by trichostatin A. *J. Biol. Chem.*, **265**, 17174–17179.
24. Marks,P.A. and Xu,W.-S. (2009) Histone deacetylase inhibitors: potential in cancer therapy. *J. Cell. Biochem.*, **107**, 600–608.
25. Lawson,B., Brewer,J.W. and Hendershot,L.M. (1998) Geldanamycin, an hsp90/GRP94-binding drug, induces increased transcription of endoplasmic reticulum (ER) chaperones via the ER stress pathway. *J. Cell. Physiol.*, **174**, 170–179.
26. Xu,L., Li,Y., Dai,Y. and Peng,J. (2018) Natural products for the treatment of type 2 diabetes mellitus: pharmacology and mechanisms. *Pharmacol. Res.*, **130**, 451–465.
27. Hebert,M.D. and Poole,A.R. (2017) Towards an understanding of regulating Cajal body activity by protein modification. *RNA Biol.*, **14**, 761–778.
28. Ho,T.T., Tran,Q.T. and Chai,C.L. (2018) The polypharmacology of natural products. *Future Med. Chem.*, **10**, 1361–1368.
29. Egan,E.D. and Collins,K. (2012) Biogenesis of telomerase ribonucleoproteins. *RNA*, **18**, 1747–1759.
30. Praveen,K., Wen,Y., Gray,K.M., Noto,J.J., Patlolla,A.R., Van Duyne,G.D. and Matera,A.G. (2014) SMA-causing missense mutations in survival motor neuron (Smn) display a wide range of phenotypes when modeled in Drosophila. *PLoS Genet*, **10**, e1004489.
31. Göckler,N., Jofre,G., Papadopoulos,C., Soppa,U., Tejedor,F.J. and Becker,W. (2009) Harmine specifically inhibits protein kinase DYRK1A and interferes with neurite formation. *FEBS J.*, **276**, 6324–6337.
32. Messner,C.B., Demichev,V., Wendisch,D., Michalick,L., White,M., Freiwald,A., Textoris-Taube,K., Vernardis,S.I., Egger,A.-S., Kreidl,M. *et al.* (2020) Ultra-high-throughput clinical proteomics reveals classifiers of COVID-19 infection. *Cell Syst*, **11**, 11–24.
33. Messner,C.B., Demichev,V., Bloomfield,N., Yu,J.S.L., White,M., Kreidl,M., Egger,A.-S., Freiwald,A., Ivosev,G., Wasim,F. *et al.* (2021) Ultra-fast proteomics with Scanning SWATH. *Nat. Biotechnol.*, **39**, 846–854.
34. Perez-Riverol,Y., Bai,J., Bandla,C., García-Seisdedos,D., Hewapathirana,S., Kamatchinathan,S., Kundu,D.J., Prakash,A., Frericks-Zipper,A., Eisenacher,M. *et al.* (2022) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.*, **50**, D543–D552.