# Supplementary Materials

## Figures

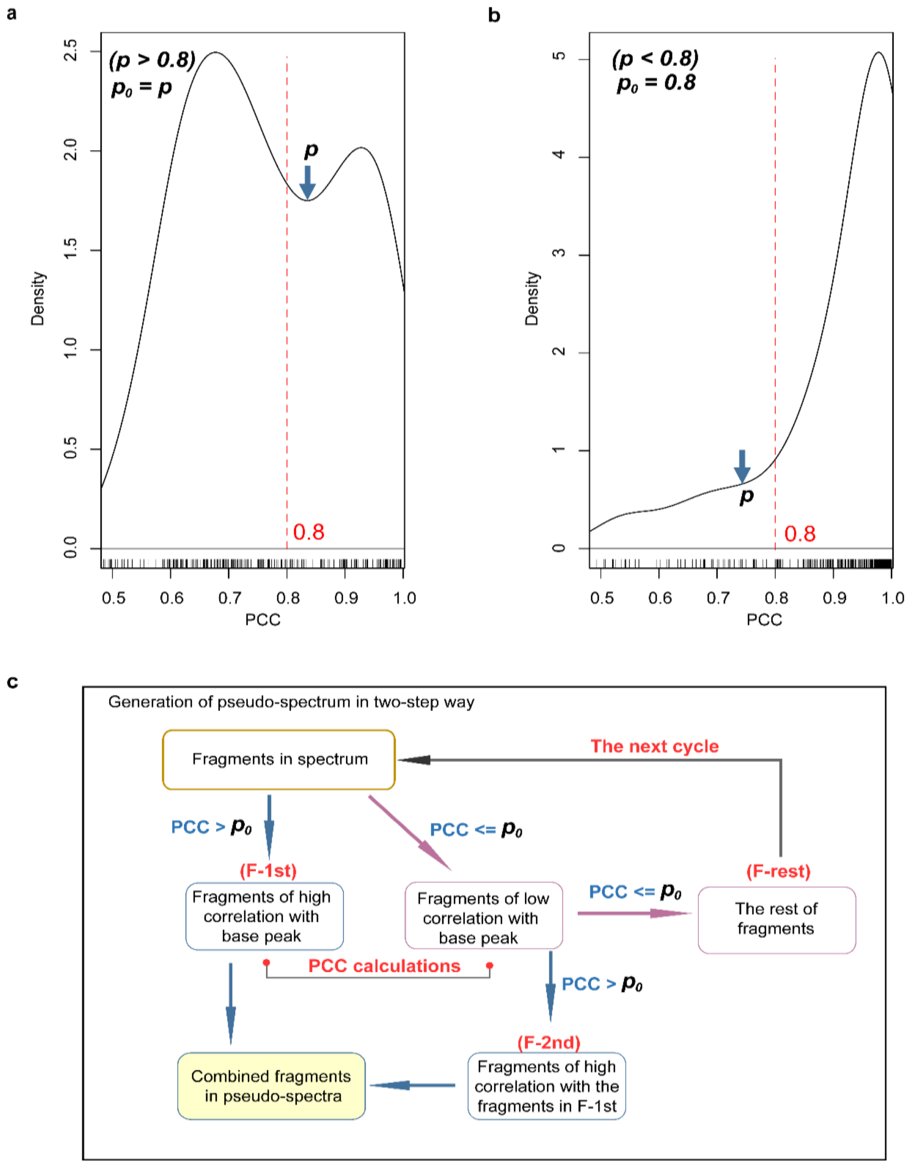


**Fig. S1** One-dimensional density kernel estimation of Pearson correlation coefficient (PCC) of the intensity profiles between the base peak and the other fragments (base peak-fragment correlations). Based on the curve of density, DIA-MS2pep determines *p* as the boundary between the Gaussian distributions. Thereby, the threshold of the PCC value ($p_0$) is set as MIN{p, 0.8}. Two typical conditions are illustrated (Panel **a** and **b**). **c** Workflow of generating pseudo-spectrum in two-step way using correlation of base peak-fragment and intra-fragment

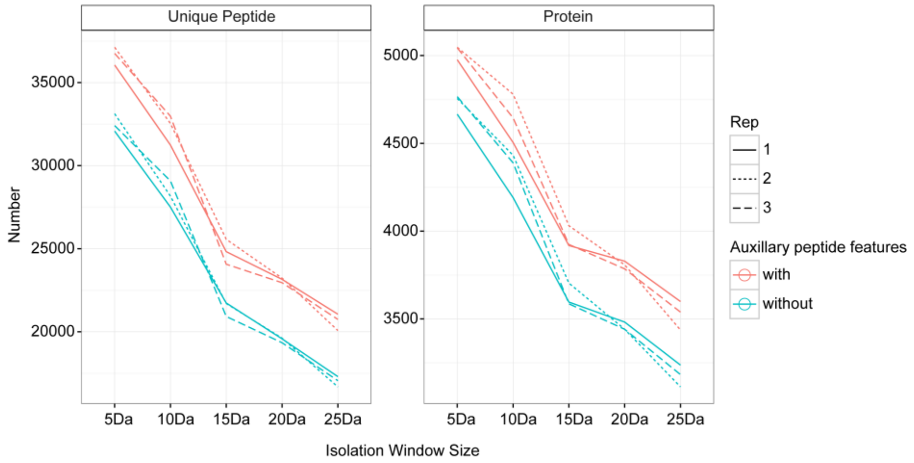


**Fig. S2** The comparison of the number of unique peptides and proteins identified from the HeLa_DIA dataset using Percolator with or without auxiliary peptide features ($q$-value < 0.01)

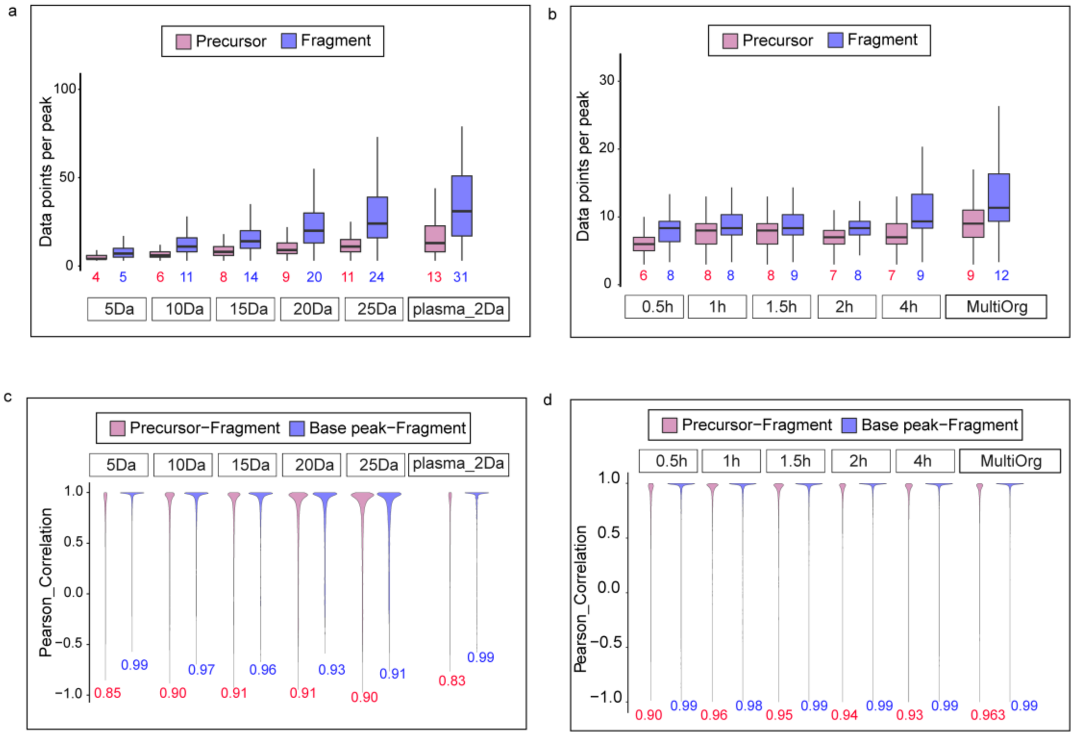


**Fig. S3** Rationale of spectrum self-demultiplexing. **a, b** Box plots of data points per peak of the precursor and fragment levels. **c, d** The distribution of the Pearson correlation coefficient of the precursor-fragment correlation and the base peak-fragment correlation. The HeLa_DIA (5Da, 10Da, 15Da, 20Da and 25Da), Plasma_GPF_DIA (plasma_2Da), HeLa_gradient_DIA (0.5, 1, 1.5, 2 and 4 h) and MultiOrg_DIA (MultiOrg) datasets (Table S1) were used as testing data

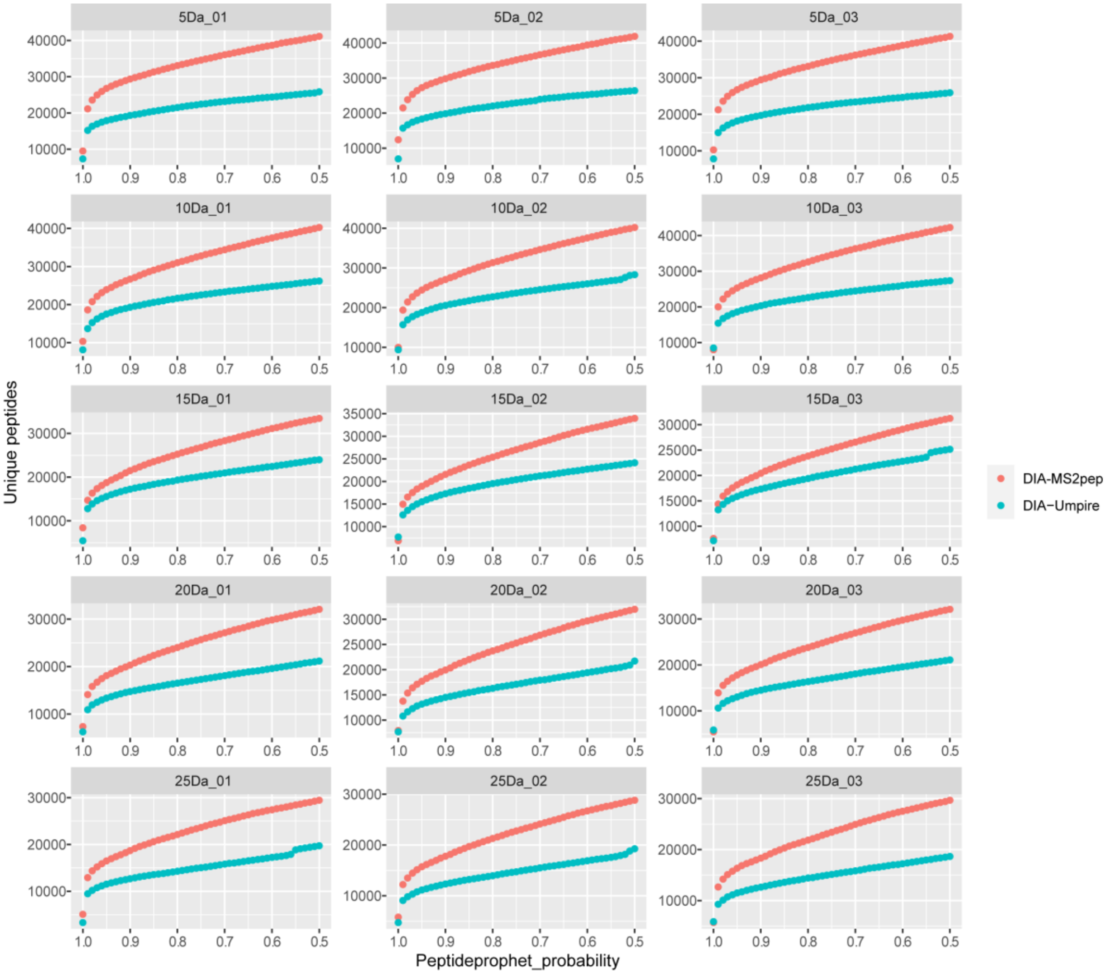


**Fig. S4** The number of unique peptides as a function of peptide probability reported by PeptideProphet in HeLa_DIA dataset

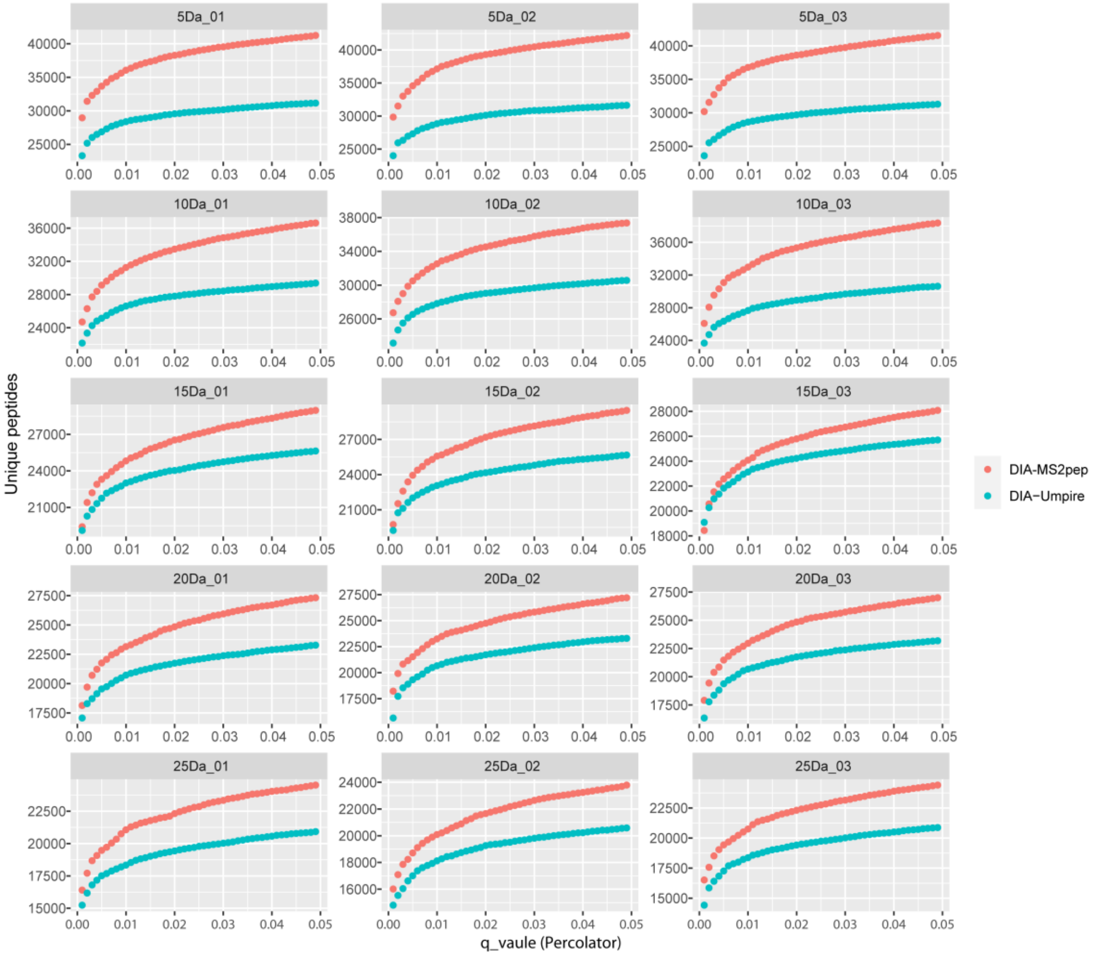


**Fig. S5** The number of unique peptides as a function of q-value reported by Percolator in HeLa_DIA dataset

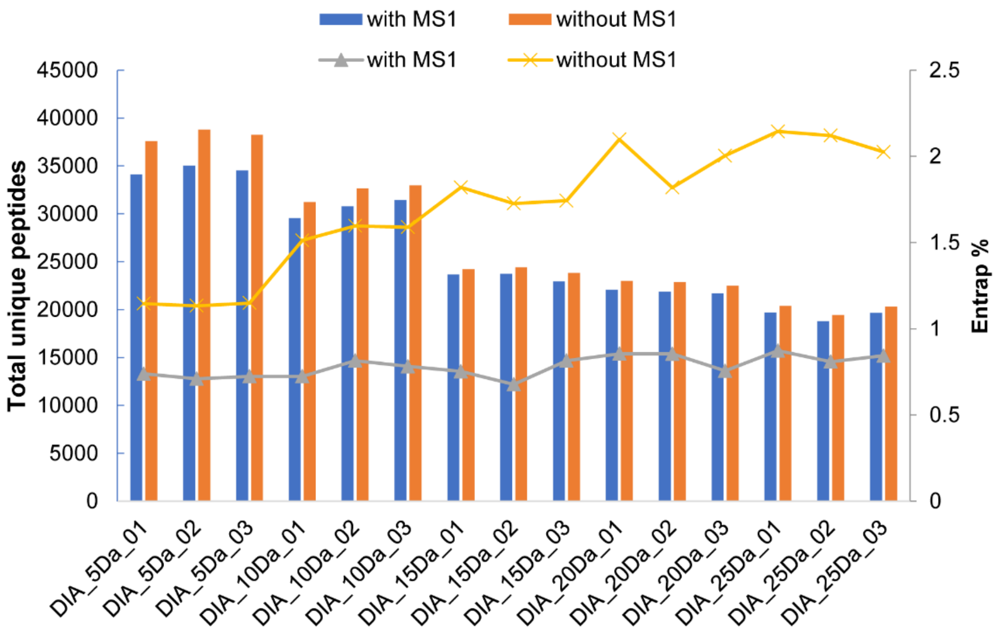


**Fig. S6** Comparison of the number of unique peptides from HeLa_DIA dataset reported by DIA-MS2pep with or without MS1 validation using entrapment method. We use presence or absence of MS1 signal as an auxiliary feature in FDR control

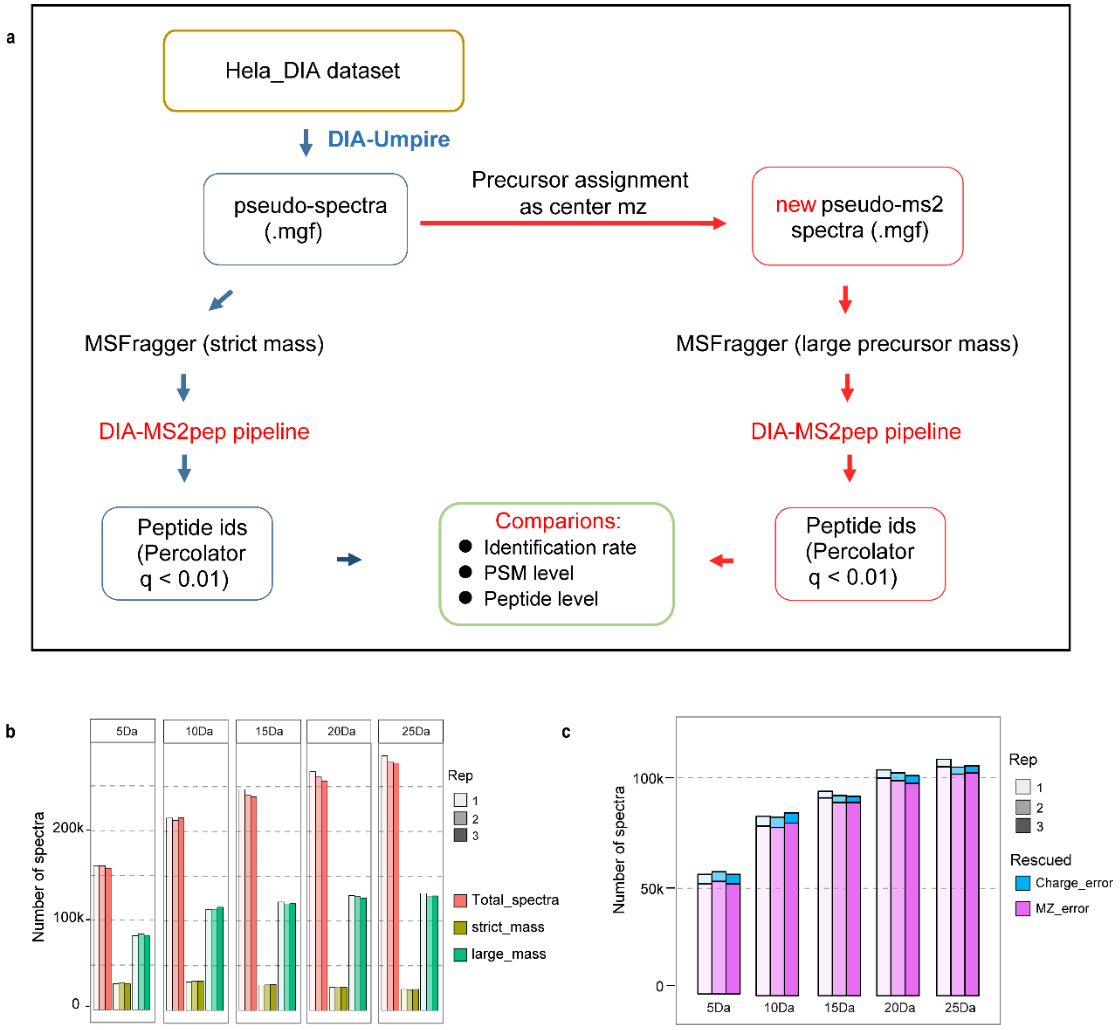


**Fig. S7** Simulation experiment using pseudo-spectra generated by DIA-Umpire from HeLa DIA dataset. **a** DIA-Umpire pseudo-spectra are searched in two ways: (1) either searched by MSFragger with strict precursor mass; (2) first modified by replacing original precursor mass as the center *m*/*z* of isolation window, and then searched by MSFragger with large precursor mass. To facilitate the fair comparison, the results from two search modes are submitted to DIA-MS2pep pipeline for data refinement, and FDR validation with Percolator (*q*-value less than 0.01). **b** Pseudo-spectra generated by DIA-Umpire from HeLa_DIA dataset is searched with strict precursor mass or large precursor mass as shown in Figure S6a. At a 1% FDR of PSM level, significantly more pseudo-spectra are identified by large precursor mass-based database search (large_mass, yellow bar) than by strict precursor mass-based database search (strict_mass, green bar). The number of total pseudo-spectra generated by DIA-Umpire from HeLa_DIA dataset is plotted as well (strict_mass, red bar). **c** Bar plots of the number of pseudo-spectra with incorrect precursor m/z or charge state assigned by DIA-Umpire. These spectra are not identified using strict precursor mass search but rescued by large precursor mass search

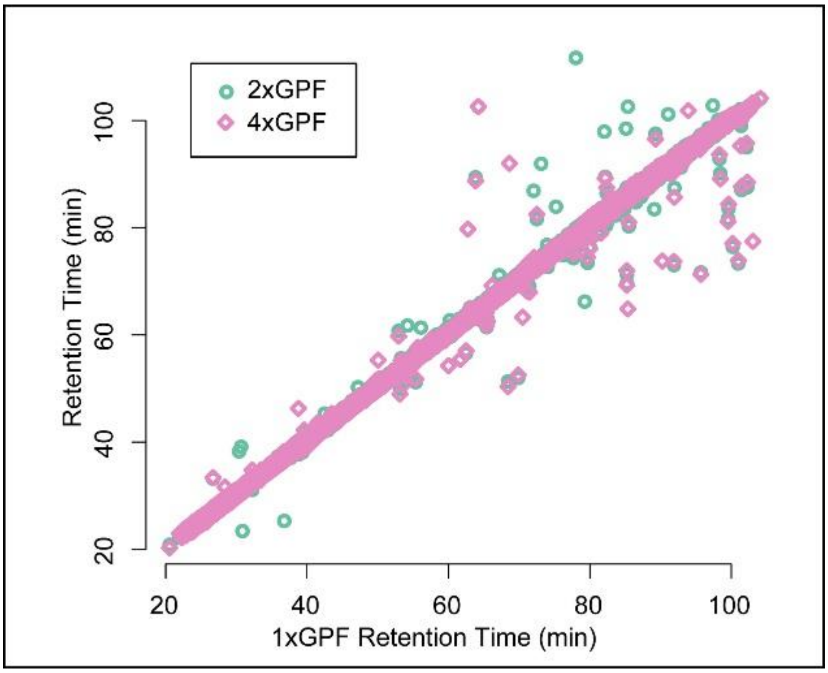


**Fig. S8** Retention time comparison of 16596 peptides commonly identified by DIA-MS2pep from 1×GPF relative to 2×GPF and 4×GPF. Peptides (0.36% and 0.31%) were identified with a discrepancy in retention time (>2.5 min) in either the 2×GPF or 4×GPF dataset compared with the 1×GPF dataset, respectively

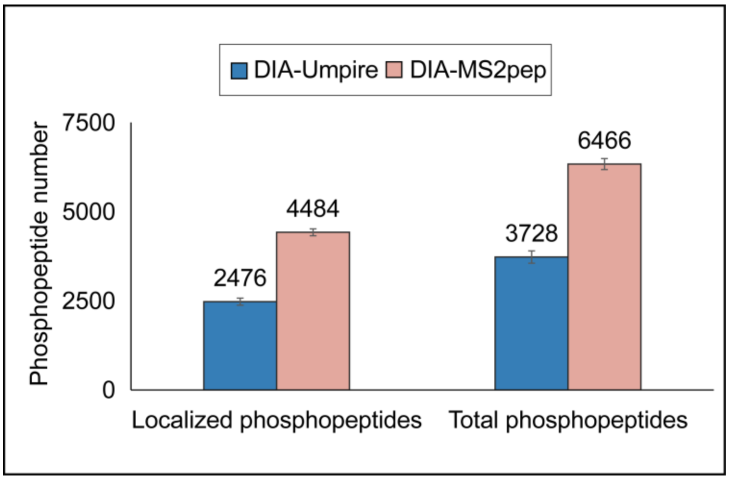


**Fig. S9** The number of phosphopeptides identified and localized from the PhosphoHeLa_DIA dataset (Searle *et al.* 2019) by DIA-Umpire or DIA-MS2pep. Phosphorylation sites with a probability of > 0.75 are required for confident localization

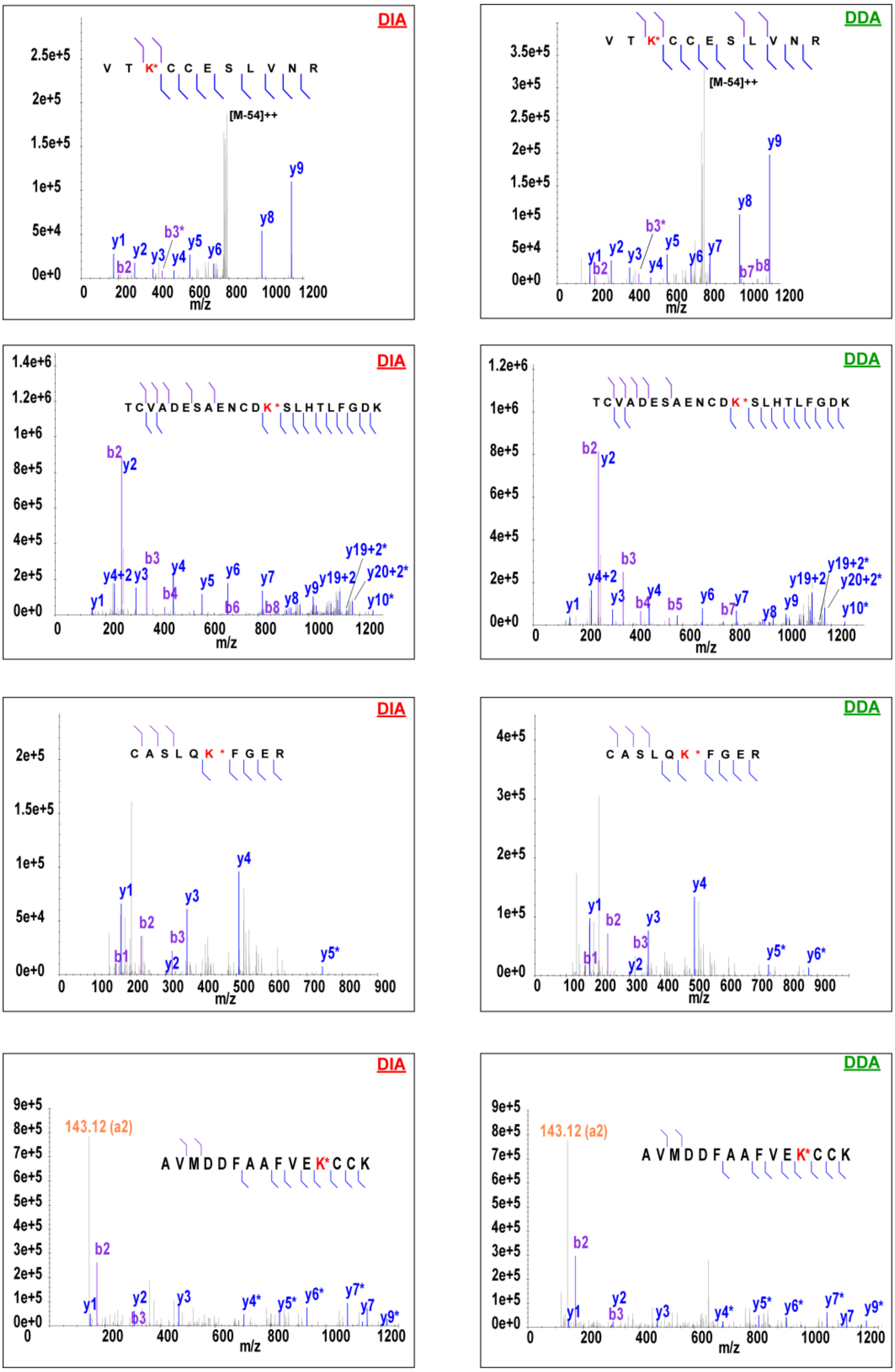


**Fig. S10** The DIA-MS2pep pseudo-spectra from Plasma_GPF_DIA dataset (left panels) vs. DDA spectra (right panels) from the sample of *in vitro* glycation experiment (Supporting Information Methods). In the spectra, b- and y-ions are denoted using purple and blue color, respectively. In addition, the neutral loss peaks of glycation ($H_6O_3$, -54 Da) are also denoted with b* and y* ions

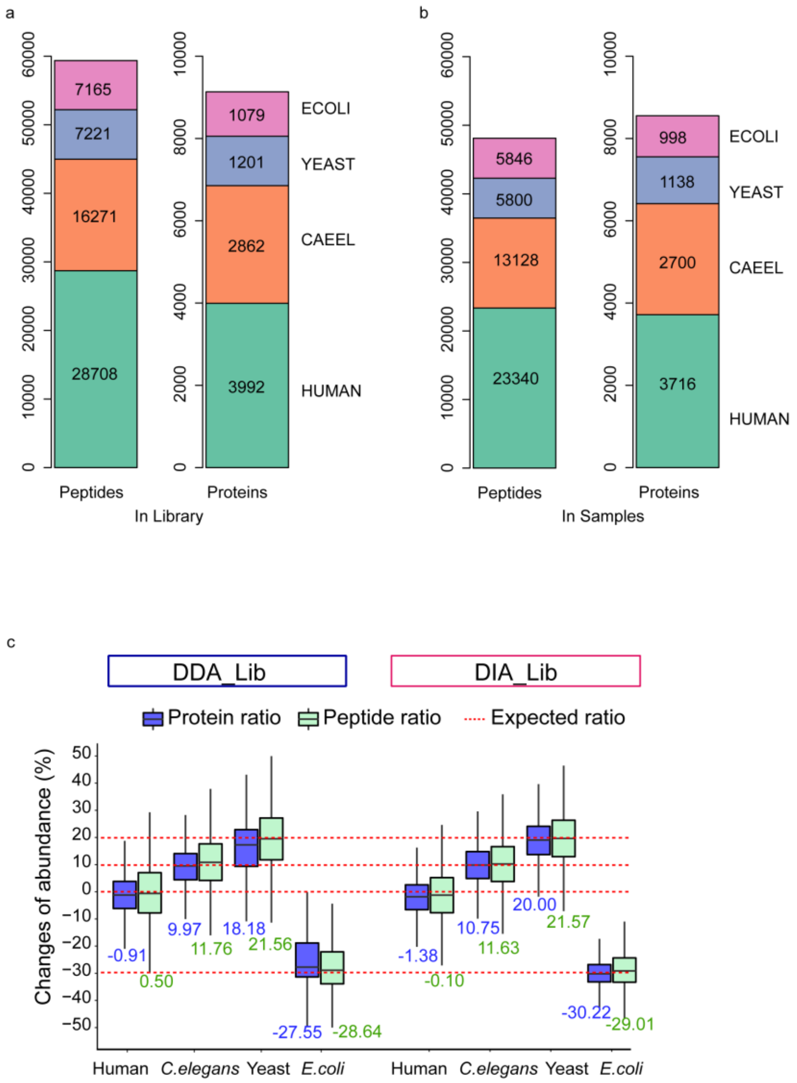


**Fig. S11** Data-specific library of MultiOrg_DIA data (Bruderer *et al.* 2017). **a** The number of peptides and proteins identified from the MultiOrg_DIA dataset by DIA-MS2pep. **b** The number of peptides and proteins quantified by EncyclopeDIA (Searle *et al.* 2018) from the MultiOrg_DIA dataset. **c** Box plot of the percent changes in abundance at the peptide and protein levels quantified from two mixed proteome samples using either the DDA library (DDA_Lib) or DIA data-specific library (DIA_Lib) built with DIA-MS2pep. The average percent change of each peptide or protein was calculated from three replicates of each sample. The median percent change of each organism is labeled. The boxes indicate the interquartile ranges (IQRs), and the whiskers indicate 1.5 × IQR values; outliers are not shown. The dashed red lines represent the expected percent changes for each organism

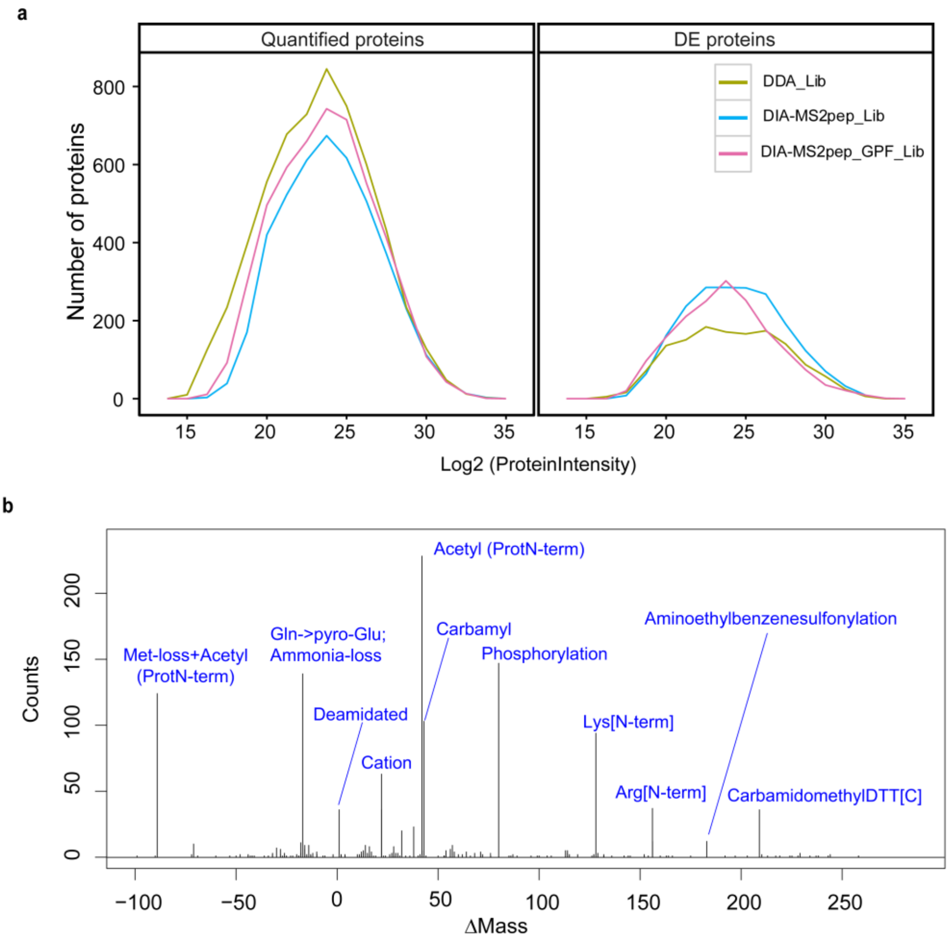


**Fig. S12** Quantitative analysis of the HeLa_Serum_DIA dataset (Searle *et al.* 2018) using five different spectral libraries: DDA_Lib, MS2pep_Lib, MS2pep_GPF_Lib. **a** The count distribution of quantified and DE proteins as a function of their expression quantities (protein intensity). **b** The mass difference profiles of 1,683 peptides identified with putative modifications by DIA-MS2pep. The representative modifications detected are labeled (blue)

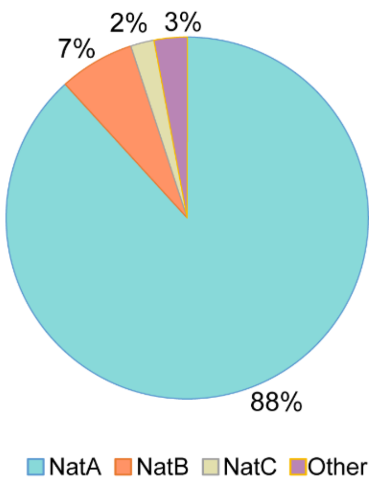


**Fig. S13** The proportional distribution of protein N-terminal acetylation is classified based on the substrates of different types of N-terminal acetyltransferases (Nats): NatA (A, C, G, S, T and V), NatB (Met-E, Met-D and Met-N) and NatC (Met-L, and Met-I) (Arnesen *et al*. 2009)

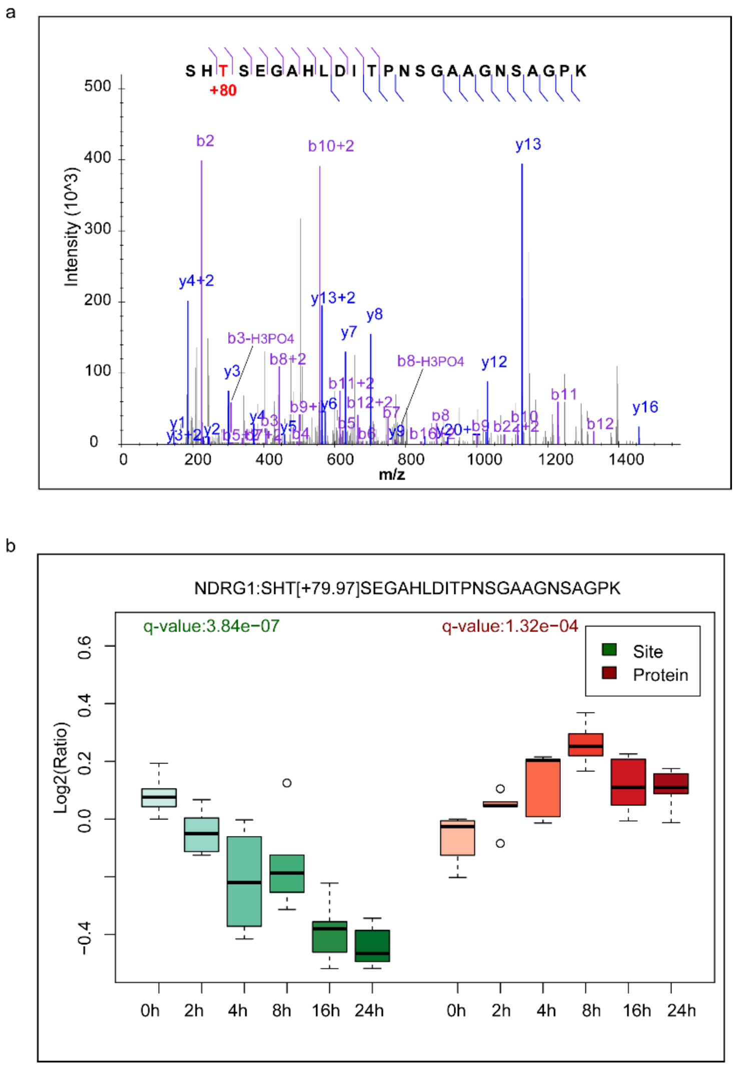


**Fig. S14** One phosphopeptide on NDRG1 identified by DIA-MS2pep (SHT[+79.97]SEGAHLDITPNSGAAGNSAGPK) from the HeLa_Serum_DIA dataset. **a** The pseudo-MS/MS spectrum denoted with b- and y-ions. **b** Box plots of the quantitative results at the phosphopeptide and protein levels. The *q*-value indicates the significance of the quantitative changes over time reported by EdgeR (Lund *et al.* 2012)

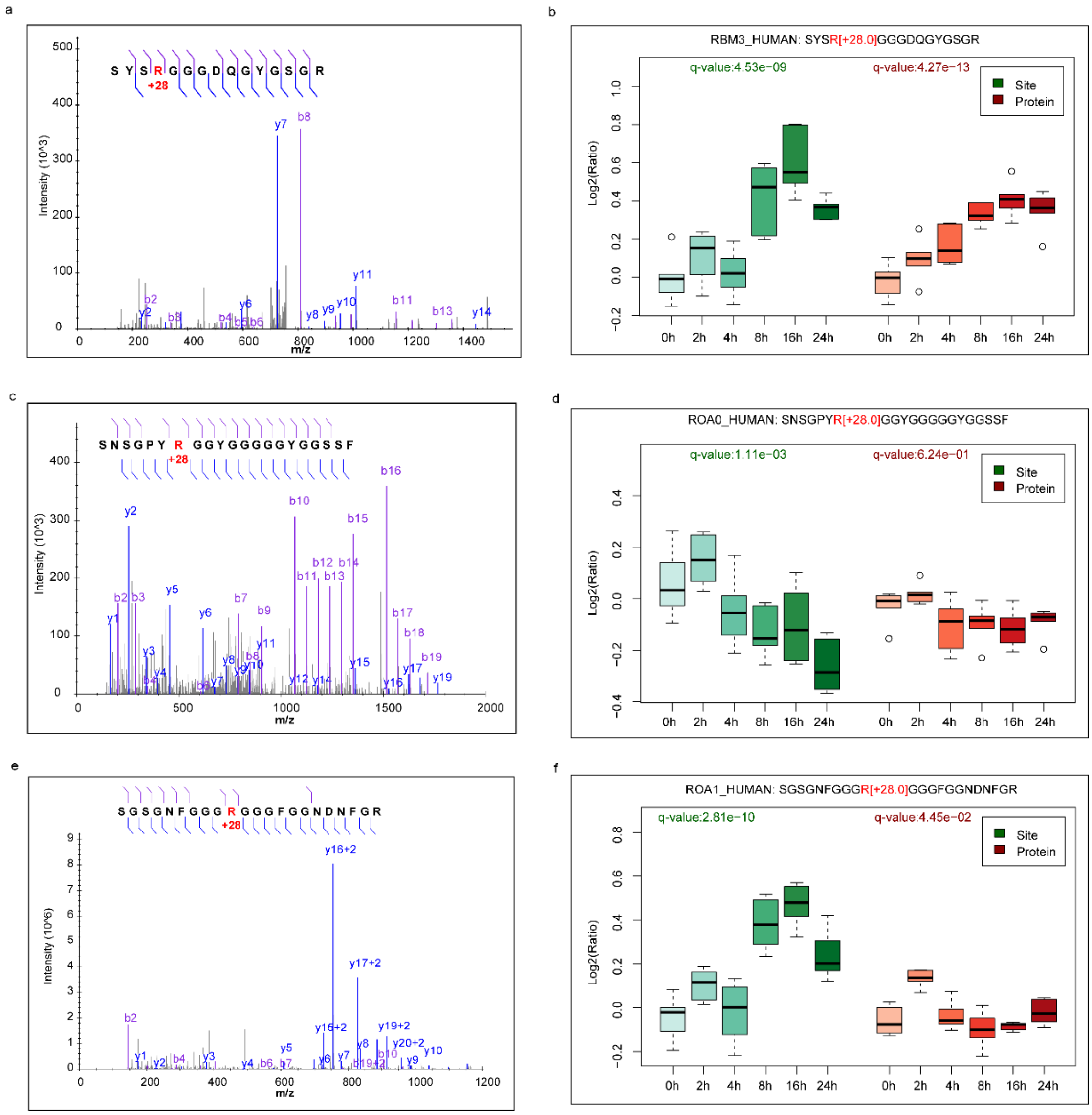


**Fig. S15** Three arginine dimethylation sites from three proteins are identified: HNRNPA0 (SNSGPYR[+28.0313]GGYGGGGGYGGSSF), HNRNPA1 (SGSGNFGGGR[+28.0313]GGGFGGNDNFGR), and RBM3 (SYSR[+28.0313]GGGDQGYGSGR). The pseudo-MS/MS spectra were denoted with b- and y-ions. Box plots displaying the quantitative results of dimethylated peptides and their corresponding proteins. The *q*-value indicates the significance of the quantitative changes over time reported by EdgeR (Lund *et al.* 2012)

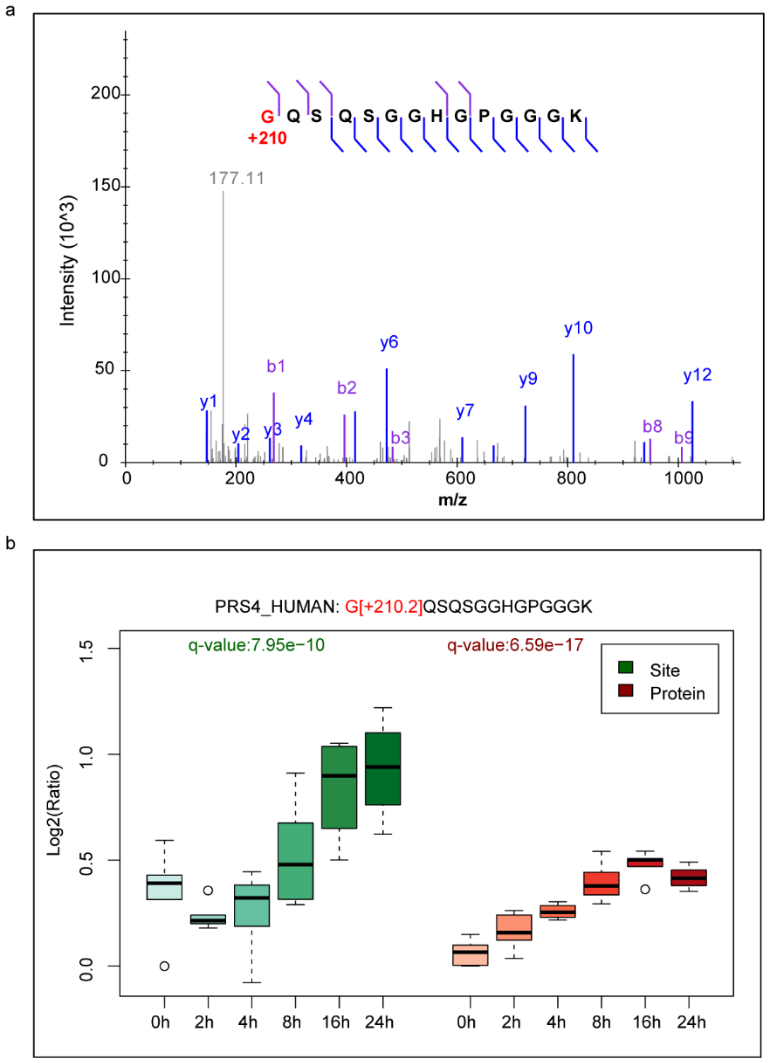


**Fig. S16** One myristoyl peptide G[+210.2]QSQSGGHGPGGGK from protein PSMC1 (PRS4_HUMAN, 26S proteasome regulatory subunit 4) is confidently identified. **a** Pseudo-MS/MS spectrum denoted with b- and y-ions. **b** Box plots of the quantitative results of the myristoyl peptide and PSMC1 at the protein level. The *q*-value indicates the significance of the quantitative changes over time reported by EdgeR (Lund *et al.* 2012)

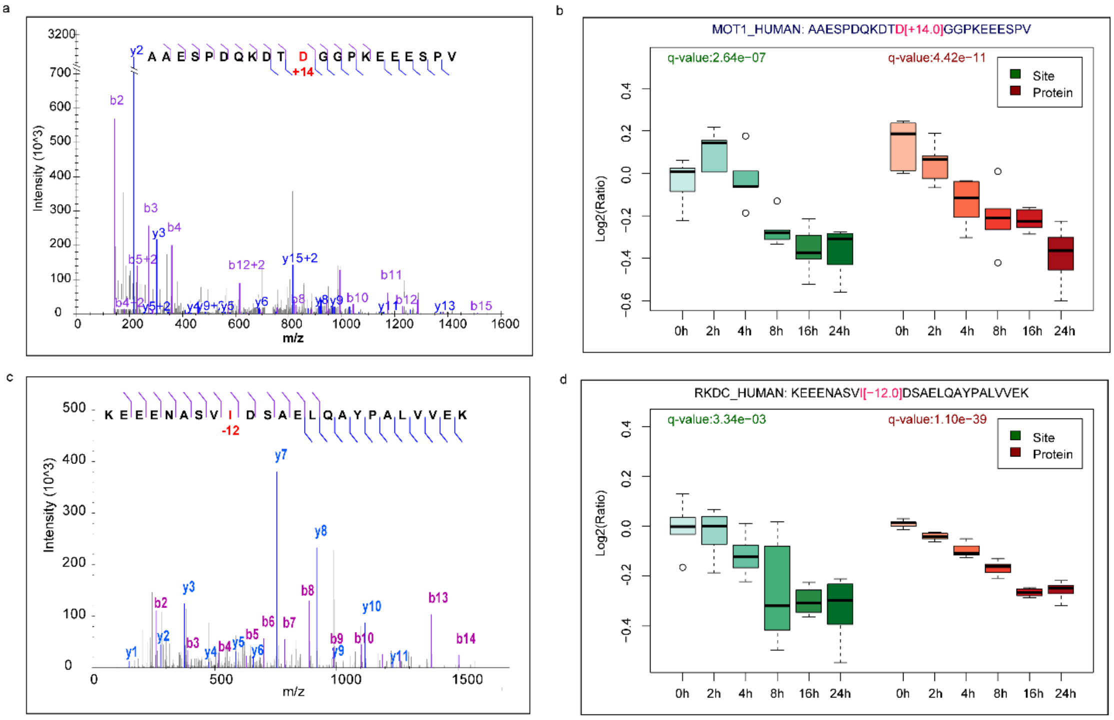


**Fig. S17** Random display of two peptide variants identified from the HeLa_Serum_DIA dataset. Both pseudo-MS/MS spectra are denoted with b- and y-ions. Box plots displaying the quantitative results of the peptide variants and their corresponding proteins. The *q*-value indicates the significance of the quantitative changes over time reported by EdgeR (Lund *et al.* 2012)

# Tables

There are five supplementary tables, all of them were uploaded on https://github.com/SS2proteome/DIA-MS2pep/tree/master/Paper/SupplementTables.

# Methods

## Comparison between DIA-MS2pep and DIA-Umpire for the HeLa_DIA dataset

To compare the performance of spectrum demultiplexing between DIA-MS2pep and DIA-Umpire, the pseudo-spectra generated by the two methods are processed with two FDR estimator-based workflows.

*PeptideProphe*

After finishing the pipeline of DIA-MS2pep, we also generate the similar pseudo-spectra as those generated by DIA-Umpire, which contain the accurate precursor information from MS1 scan. The pseudo-spectra from the two methods are converted to mzML files using MSconvert (part of ProteoWizard 28, v3.0.9974), and then the remaining data are processed with PeptideProphet-based workflow in FragPipe software (v15.0); version information: MSFragger v2.4 and Philosopher v3.4.13. Peptideprophet is run with the option “--decoyprobs --ppm --accmass --nonparam --expectscore --decoy REV_”, where “REV_” is the prefix of the decoy protein. Finally, the search data are filtered and reported by executing the command: “philosopher.exe filter --sequential --razor --mapmods --prot 0.01 --tag REV_ --pepxml”.

*Percolator*

For DIA-Umpire, the pseudo-spectra are first searched with MSFragger v2.4 and all the search results are stored in pepXML files. Using an in-house script, these pepXML files are converted to PIN-format files, which contain the same peptide features just as those in PIN-format files generated by DIA-MS2pep pipeline. Finally, PIN-format files from two methods are validated using Percolator (v3.02.1).

## Simulation experiment for the HeLa_DIA dataset

To test which type of database search is the more efficient method to interpret the pseudo-spectra from DIA data, strict precursor mass or large precursor mass, we design the simulation experiment as shown in Fig. S7a. The dataset is first processed by DIA-Umpire, and then the pseudo-spectra generated by DIA-Umpire from HeLa_DIA dataset are processed in two different ways. (1) Database search by MSFragger using strict precursor mass, and the resulting pepXML files are processed through an in-house script to generate PIN-format files, which are then validated by Percolator. (2) the original pseudo-spectra are first modified as DIA-MS2pep-like pseudo-spectra by replacing the precursor assigned by DIA-Umpire as the center *m*/*z* of the isolation window, and then the new pseudo-spectra are searched by MSFragger with large precursor mass followed by DIA-MS2pep pipeline. The comparison of identifications at PSM and peptide levels are based on the q-value reported by Percolator.

### *In vitro* glycation of albumin and MS analysis

Glycated albumin was prepared as described in the previous study (Soudahome *et al.* 2018) with slight modification. In brief, fatty acid–free human serum albumin (HSA) (Sigma, A1887) solutions were prepared in PBS (pH 7.4) at a concentration of 100 mg/mL. HSA solution was incubated at 37 °C with glucose (90 mg/mL) for three weeks. Havested HSA protein was first separated using one-dimensional SDS-PAGE, the HSA band was manually excised. In-gel digestion procedure was followed as the previous study (Li 2018). The tryptic peptides were dried using a SpeedVac and stored at –20 °C for further MS analysis, which was performed with Orbitrap Exploris 480 mass spectrometer (Thermo Scientific) in data-dependent acquisition mode. The MS raw data was analyzed with pFind v3.0 against HSA protein sequence with searching parameters as following: Trypsin as enzyme; up to two missed cleavages; mass tolerance of precursor and fragment were set as 20 ppm; cysteine carbamidomethylation as fixed modifications; methionine oxidation and glycation on R/K chosen as variable modifications.

### Mascot with error tolerance search for Plasma_GPF_DIA dataset

Using in-house script, we first generate the pseudo-spectra, that are identified as confident peptide hits with FDR < 0.01 estimated by Percolator. All these spectra are assigned with accurate precursor mz and charge state, then submitted to Mascot. The main search parameters are as following: Trypsin as enzyme; two missed cleavages; Carbamidomethyl (C) as fixed modification; Oxidation (M) as variable modification; precursor mass tolerance of 2 ppm; fragment mass tolerance of 20 ppm; error tolerance search enabled. The search results were filtered by significance threshold less than 0.05.

# The instructions of DIA-MS2pep

## System requirement

DIA-MS2pep is an open-source tool well-compatible to Linux system. For Windows system, the setup of a Unix environment (Cygwin) is required (The installation process is automated and not complicated).

Before the start of DIA-MS2pep, make sure that your system is complied with the following requirements:

- Perl programming language (Version 5) with required modules: “Math”, “MIME”, “Statistics”, and “Parallel::ForkManager” *
- msconvert (ProteoWizard, Version 3.0.9974 or above)
- MSFragger (Version 2.4, or above)
- Percolator (Version 3.02.1)

* Installing perl modules from the Comprehensive Perl Archive Network (CPAN).

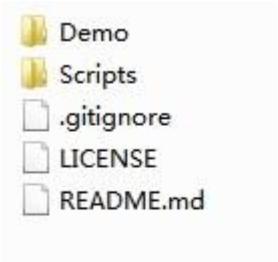
## DIA-MS2pep software

DIA-MS2pep comprises of four main components:

1. *DIA/SWATH_pesudo_MS2*: generation of pseudo-spectra from DIA data, where we provide two scripts for Obitrap and TripleTOF data, respectively.
2. *MSFragger_runner*: implementation of MSFragger to perform large precursor mass database search.
3. *DIA/SWATH_data_refinement*: verification of precursor evidence, searching for modified forms and computation of auxiliary peptide scores
4. *Percolator_runner*: implementation of Percolator to validate the peptide hits at PSM, peptide and protein level.

## How to run DIA-MS2pep

1. Download the DIA-MS2pep from Github website, and decompress the zipped file.



Demo
Scripts
.gitignore
LICENSE
README.md

“Demo” folder contains three files: DIA_MS2pep.config, DIA-MS2pep_run.sh, and Description. DIA_MS2pep.config contains the parameters setting for one DIA file (HeLa1ug_QC_Middle_DIA_5Da_150226_01.raw) from HeLa_DIA dataset, which can be downloaded from ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2016/06/ PXD003179.

“Scripts” folder contains all the codes of DIA-MS2pep.

2. Preparation of the files required

(1) .mzML file

msconvert-SIMMS1.config (Scripts/DataConvert/)

*>msconvert -c msconvert.config HeLa1ug_QC_Middle_DIA_5Da_150226_01.raw*

(2) .mgf and .ms1 file

RAW.mzML.parser.pl (Scripts/DataConvert/)

*>RAW.mzML.parser.pl mgf HeLa1ug_QC_Middle_DIA_5Da_150226_01.mzML*

*>RAW.mzML.parser.pl mgf HeLa1ug_QC_Middle_DIA_5Da_150226_01.mzML*

(3) Protein database

uniprot-proteome_Human_20200310_no_Fragment_iRTpeptides.decoy_target.fasta, which contains combined target and decoy protein sequences.

3. Configuration setting of parameter file *DIA_MS2pep.config* (Demo/)

*“MS2pepCodedir”*: Path to the folder of DIA-MS2pep

*“filelist”*: A list of filenames to be analyzed by DIA-MSepep

*“ms1ppm”*: Precursor mass tolerance

*“ms2ppm”*: Fragment mass tolerance

*“fasta”*: Path to the protein database file in FASTA format, combined target and decoy protein sequences

*“MSFragger_search_engine”*: Path to MSFragger search tool, like MSFragger-2.4.jar

*“MSFragger_search_params”*: Path to the parameter file of MSFragger

*“PTM”*: Set as 1 for the sample that is enriched with modified peptides, like phosphorylated peptides; Default as 0

*“ptm_search”*: Set as 1 for wide precursor mass tolerance [-150, 500] is used when data-searched with MSFragger; Default as 0

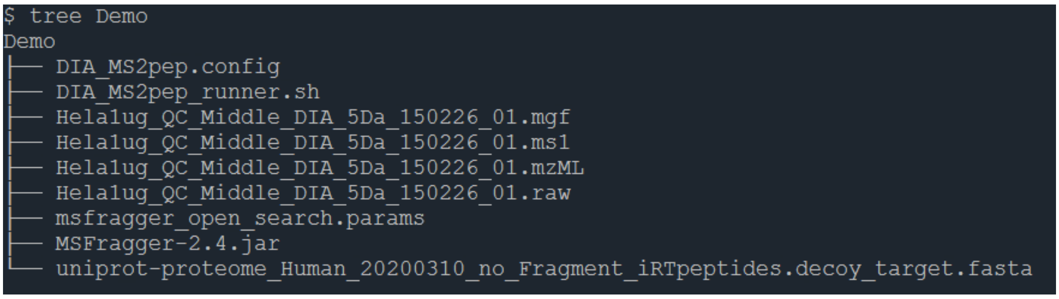
“*max_processes*”: Set the maximum number of threads

“*datatype*”: Set “DIA” or “SWATH” to indicate the type of data collected by different instrument

4. Implement of DIA-MS2pep pipeline via *DIA_MS2pep_runner.sh* (Demo/)

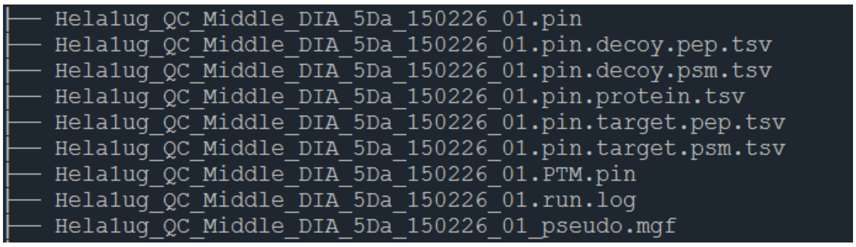
*>sh DIA_MS2pep_runner.sh DIA_MS2pep.config HeLa1ug_QC_Middle_DIA_5Da_150226_01*

** before run the above command, please check out the required files as below:*



```
$ tree Demo
Demo
├── DIA_MS2pep.config
├── DIA_MS2pep_runner.sh
├── Hela1ug_QC_Middle_DIA_5Da_150226_01.mgf
├── Hela1ug_QC_Middle_DIA_5Da_150226_01.ms1
├── Hela1ug_QC_Middle_DIA_5Da_150226_01.mzML
├── Hela1ug_QC_Middle_DIA_5Da_150226_01.raw
├── msfragger_open_search.params
├── MSFragger-2.4.jar
└── uniprot-proteome_Human_20200310_no_Fragment_iRTpeptides.decoy_target.fasta
```

Output files will be generated in folder “MS2pep” as below:



```
├── Hela1ug_QC_Middle_DIA_5Da_150226_01.pin
├── Hela1ug_QC_Middle_DIA_5Da_150226_01.pin.decoy.pep.tsv
├── Hela1ug_QC_Middle_DIA_5Da_150226_01.pin.decoy.psm.tsv
├── Hela1ug_QC_Middle_DIA_5Da_150226_01.pin.protein.tsv
├── Hela1ug_QC_Middle_DIA_5Da_150226_01.pin.target.pep.tsv
├── Hela1ug_QC_Middle_DIA_5Da_150226_01.pin.target.psm.tsv
├── Hela1ug_QC_Middle_DIA_5Da_150226_01.PTM.pin
├── Hela1ug_QC_Middle_DIA_5Da_150226_01.run.log
├── Hela1ug_QC_Middle_DIA_5Da_150226_01_pseudo.mgf
```

where .pin file contains all the peptide hits of target and decoy; .psm.tsv, .pep.tsv and protein.tsv are the result files reported by Percolator at PSM, peptide and protein levels.

## References

Arnesen T, Van Damme P, Polevoda B, Helsens K, Evjenth R, Colaert N, Varhaug JE, Vandekerckhove J, Lillehaug JR, Sherman F, Gevaert K (2009) Proteomics analyses reveal the evolutionary conservation and divergence of N-terminal acetyltransferases from yeast and humans. Proc Natl Acad Sci USA 106(20):8157-8162

Bruderer R, Bernhardt OM, Gandhi T, Xuan Y, Sondermann J, Schmidt M, Gomez-Varela D, Reiter L (2017) Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. Mol Cell Proteomics 16(12):2296-2309

Li M, Du W, Zhou M, Zheng L, Song E, Hou J (2018) Proteomic analysis of insulin secretory granules in INS-1 cells by protein correlation profiling. Biophys Rep 4(6):329-338

Lund SP, Nettleton D, McCarthy DJ, Smyth GK (2012) Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. Stat Appl Genet Mol Biol 11(5):/j/sagmb.2012.11.issue-5/1544-6115.1826/1544-6115.1826.xml. doi: 10.1515/1544-6115.1826

Searle BC, Lawrence RT, MacCoss MJ, Villen J (2019) Thesaurus: quantifying phosphopeptide positional isomers. Nat Methods 16(8):703-706

Searle BC, Pino LK, Egertson JD, Ting YS, Lawrence RT, MacLean BX, Villen J, MacCoss MJ (2018) Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. Nat Commun 9(1):5128. doi: 10.1038/s41467-018-07454-w

Soudahome AG, Catan A, Giraud P, Kouao SA, Guerin-Dubourg A, Debussche X, Le Moullec N, Bourdon E, Bravo SB, Paradela-Dobarro B, Alvarez E, Meilhac O, Rondeau P, Couprie J (2018) Glycation of human serum albumin impairs binding to the glucagon-like peptide-1 analogue liraglutide. J Biol Chem 293(13):4778-4791