DIA-Umpire: comprehensive computational framework for data independent acquisition proteomics

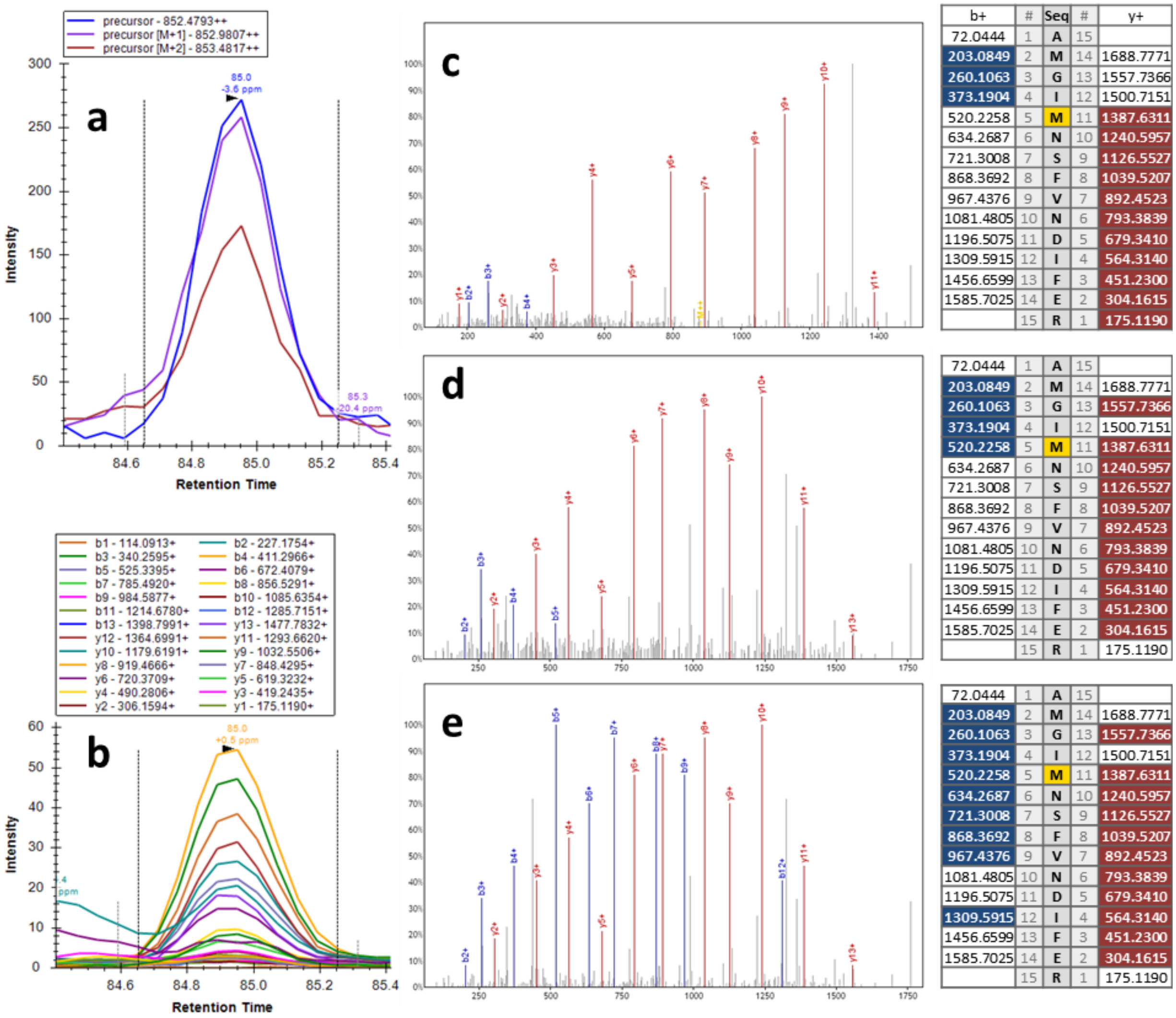
Chih-Chiang Tsou[1,2], Dmitry Avtonomov[2], Brett Larsen[3], Monika Tucholska[3], Hyungwon Choi[4]
Anne-Claude Gingras[3,5,*], Alexey I. Nesvizhskii[1,2,*]

1. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA.
2. Department of Pathology, University of Michigan, Ann Arbor, Michigan, USA.
3. Lunenfeld-Tanenbaum Research Institute, Toronto, Canada
4. Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore
5. Department of Molecular Genetics, University of Toronto, Toronto, Canada

*To whom all correspondence should be addressed:

Alexey I. Nesvizhskii nesvi@med.umich.edu; Anne-Claude Gingras gingras@lunenfeld.ca

Supplementary Figures and Tables:

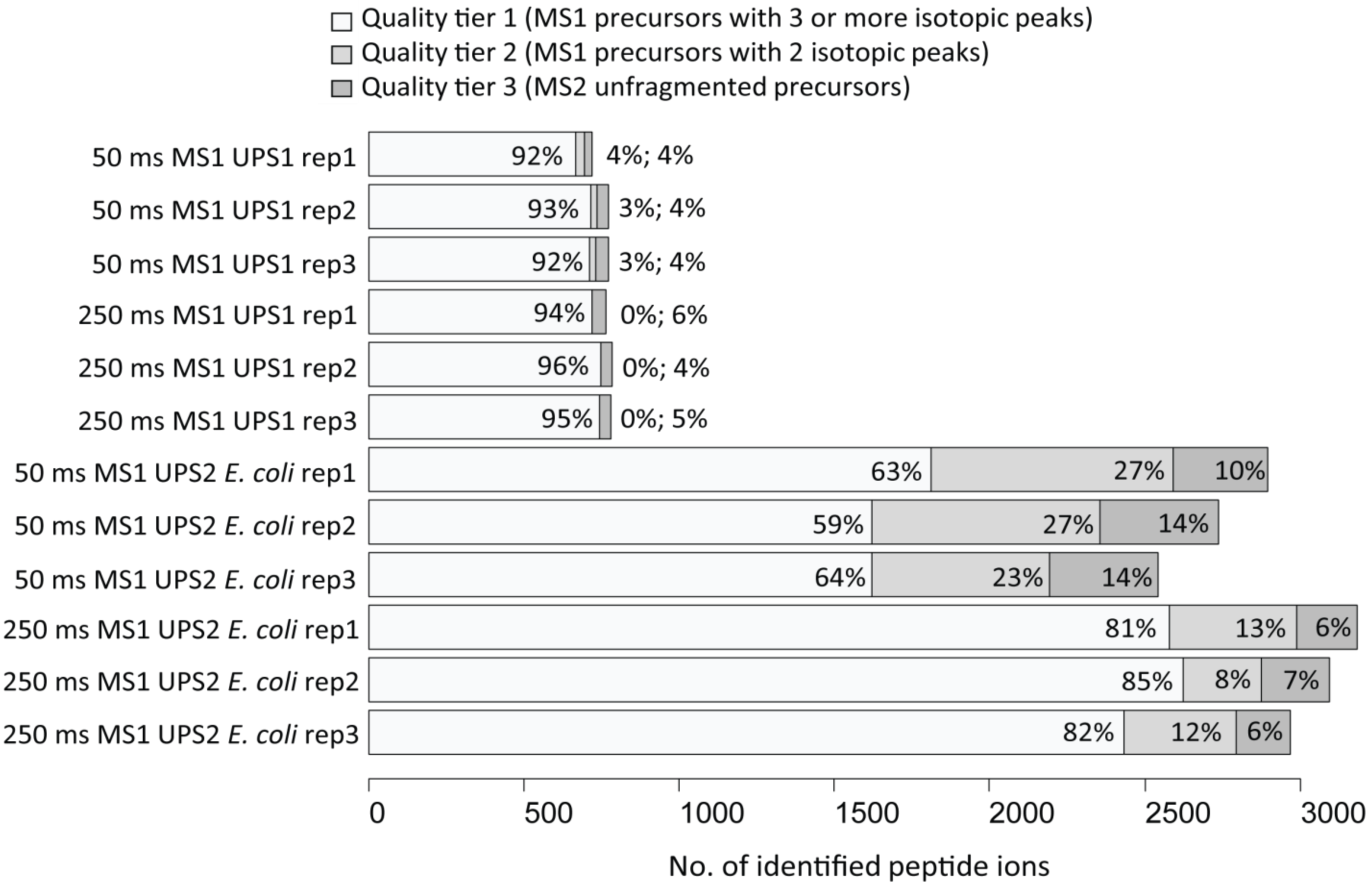


b+	#	Seq	#	y+
72.0444	1	A	15	
203.0849	2	M	14	1688.7771
260.1063	3	G	13	1557.7366
373.1904	4	I	12	1500.7151
520.2258	5	M	11	1387.6311
634.2687	6	N	10	1240.5957
721.3008	7	S	9	1126.5527
868.3692	8	F	8	1039.5207
967.4376	9	V	7	892.4523
1081.4805	10	N	6	793.3839
1196.5075	11	D	5	679.3410
1309.5915	12	I	4	564.3140
1456.6599	13	F	3	451.2300
1585.7025	14	E	2	304.1615
	15	R	1	175.1190

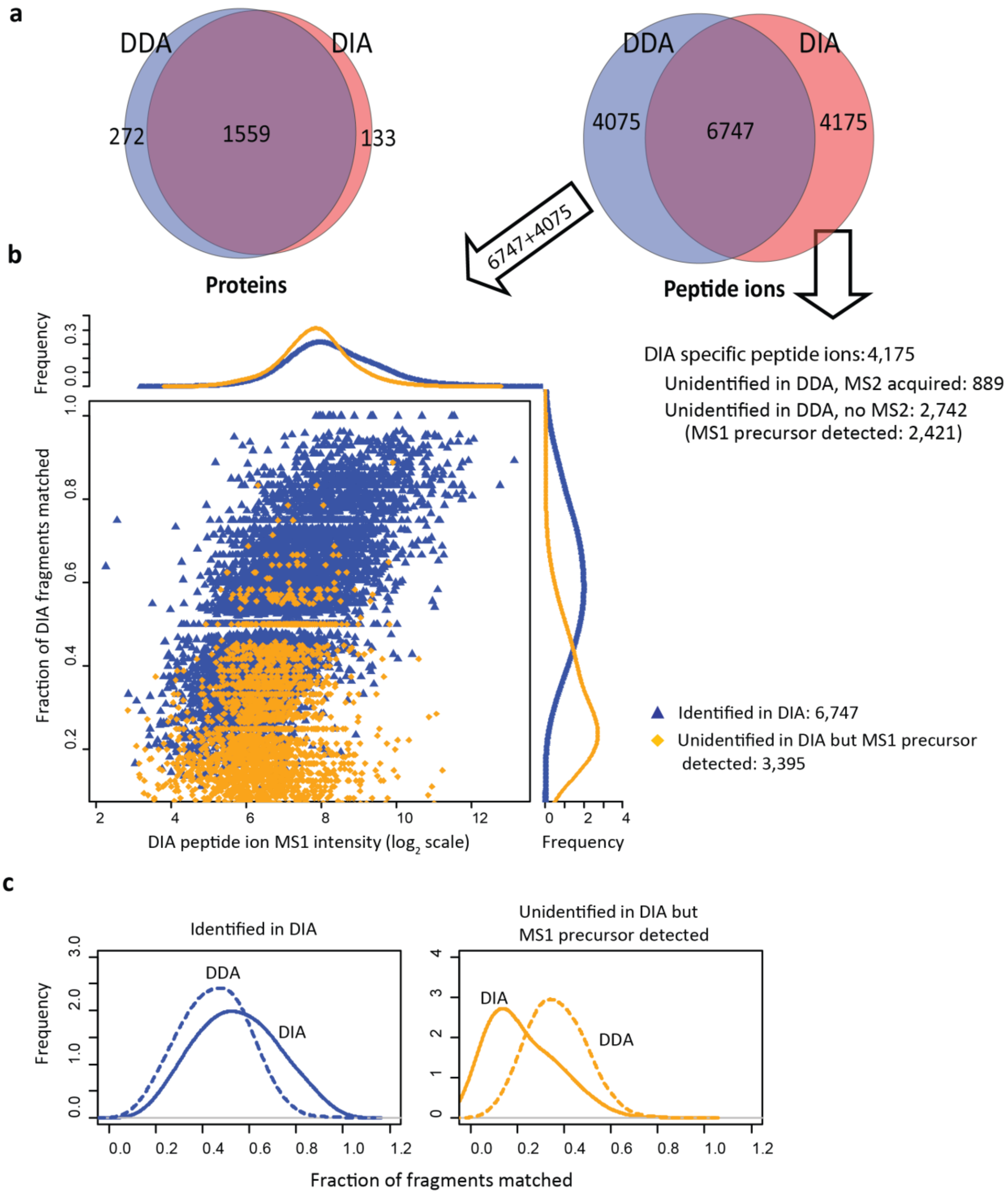
72.0444	1	A	15	
203.0849	2	M	14	1688.7771
260.1063	3	G	13	1557.7366
373.1904	4	I	12	1500.7151
520.2258	5	M	11	1387.6311
634.2687	6	N	10	1240.5957
721.3008	7	S	9	1126.5527
868.3692	8	F	8	1039.5207
967.4376	9	V	7	892.4523
1081.4805	10	N	6	793.3839
1196.5075	11	D	5	679.3410
1309.5915	12	I	4	564.3140
1456.6599	13	F	3	451.2300
1585.7025	14	E	2	304.1615
	15	R	1	175.1190

72.0444	1	A	15	
203.0849	2	M	14	1688.7771
260.1063	3	G	13	1557.7366
373.1904	4	I	12	1500.7151
520.2258	5	M	11	1387.6311
634.2687	6	N	10	1240.5957
721.3008	7	S	9	1126.5527
868.3692	8	F	8	1039.5207
967.4376	9	V	7	892.4523
1081.4805	10	N	6	793.3839
1196.5075	11	D	5	679.3410
1309.5915	12	I	4	564.3140
1456.6599	13	F	3	451.2300
1585.7025	14	E	2	304.1615
	15	R	1	175.1190

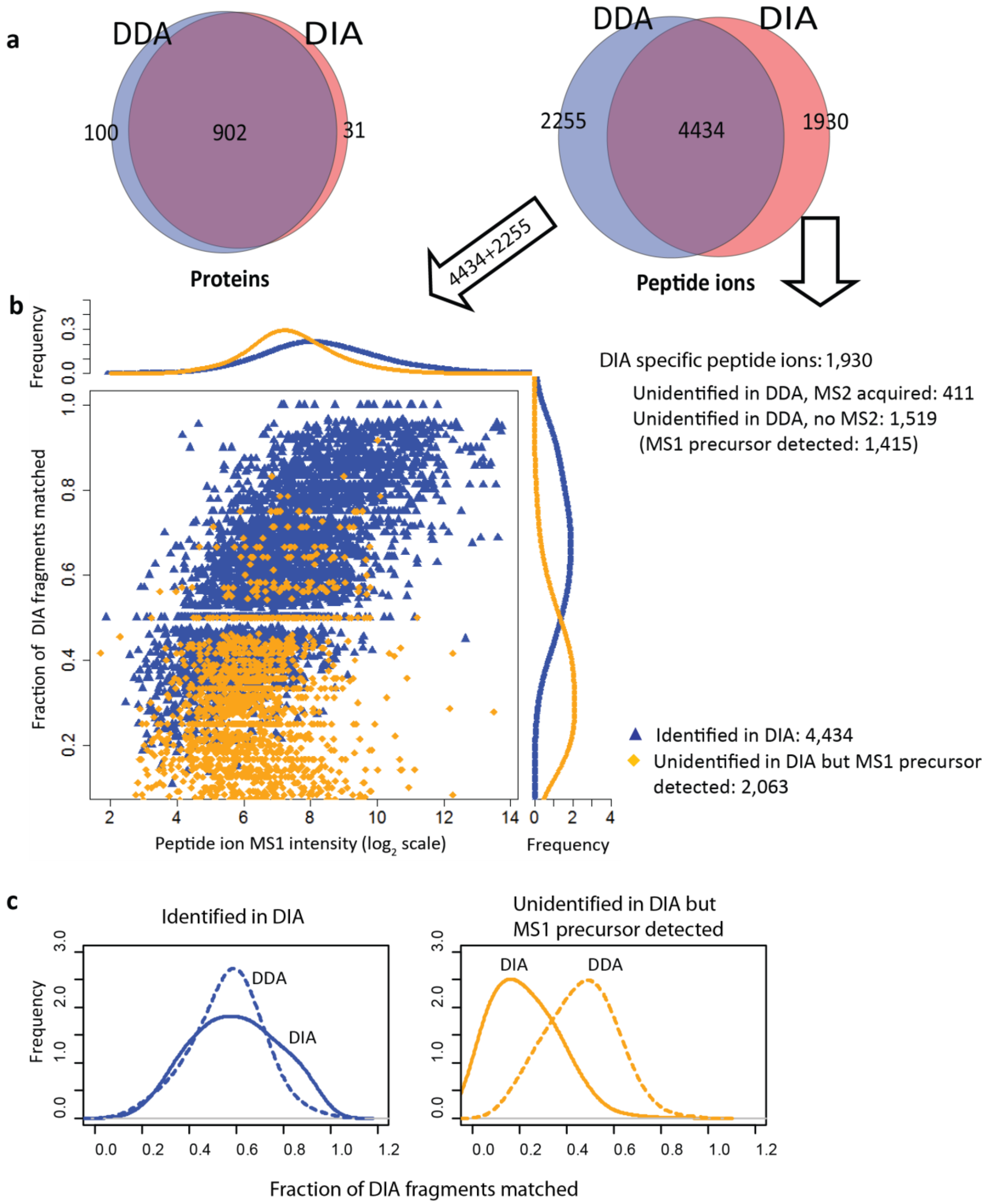
Supplementary Fig. 1. **Example of precursor peptide ion and fragment ion LC elution signals and the corresponding pseudo MS/MS spectrum generated by DIA-Umpire**. **(a)** Elution profiles for the first 3 isotopic peaks of a doubly charged precursor peptide ion AMGIM[Oxy]NSFVNDIFER extracted from MS1 data from a DIA (SWATH) run on a AB SCIEX 5600 instrument. **(b)** Elution profiles for fragments of this precursor peptide detected in the DIA MS2 data. **(c)** DDA MS/MS spectrum (from a DDA run generated on the same instrument and using the same sample) from which the same peptide was identified, with matched *b*- and *y*- ions highlighted. **(d)** Pseudo MS/MS spectrum extracted by DIA-Umpire from the DIA data (before complementary ion boosting). **(e)** Same pseudo MS/MS spectrum after complementary ion boosting. Note a larger number of *b*- ions matched in (e) compared to (d). (a) and (b) images exported from Skyline. (c), (d), and (e) are exported from TPP spectrum browser.



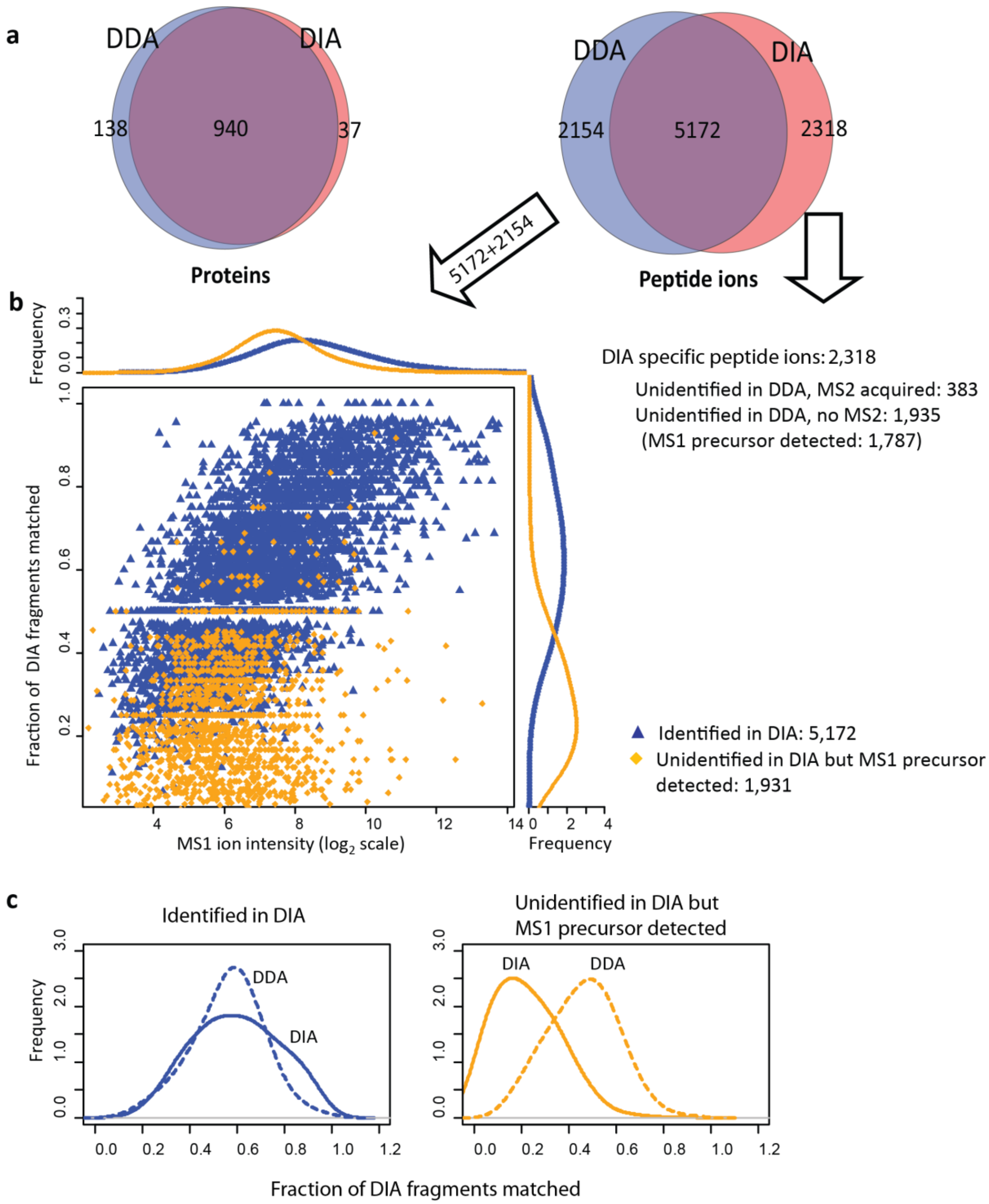
Supplementary Fig. 2 Effect of MS1 survey scan ion accumulation time on peptide identification using DIA-Umpire. Experiments to assess the identification performance of DIA-Umpire on data generated using different MS1 ion accumulation times in DIA (SWATH) analysis using AB SCIEX 5600 instrument were carried out using two samples: UPS1 proteins, and UPS2 mixture spiked in with *E. coli* background. Two settings (50 ms and 250 ms MS1 ion accumulation time) were tested. Three replicate runs were acquired for each sample/condition. The numbers shown are non-redundant contributions to the total number of peptide ion identifications in each replicate/condition from pseudo MS/MS spectra from three different quality tiers: QT = 1 (white bar), 2 (grey), and 3 (dark grey). The QT = 1 category represents pseudo MS/MS spectra that are linked to high quality MS1 precursor features (3 or more detected isotope peaks), QT = 2 represent lower abundance precursors (2 detected isotope peaks only), and QT = 3 represents unfragmented precursors detected in DIA MS2 scans. In a low complexity UPS1 sample, the dominant majority of peptide ions were identified from QT =1 spectra. Even with the using short MS1 accumulation time (50 ms), 92–93% of the peptides ions were identified from the QT = 1 spectral subset (this fraction increased slightly, to 94–96%, with the longer 250 ms accumulation time). Note that inclusion of unfragmented precursors detected in DIA MS2 data (QT = 3 subset) in the analysis contributed 4–6% of the total peptide ion identifications in UPS1 samples. In the more complex UPS2 plus *E. coli* samples, the effect of the accumulation time on the quality of MS1 signal was more pronounced. The longer DIA MS1 survey scan ion accumulation time resulted in more high quality (QT = 1) precursor peptide features detected, and thus more peptides identified from pseudo MS/MS spectra in the QT = 1 subset (81–85% for 250 ms vs. 59–64 % for 50 ms). Congruently, QT = 2 and QT = 3 spectral subsets contributed a higher percentage to the total number of peptide ion identifications when using 50 ms accumulation time setting. The overall number of identifications (from all 3 QT sets) has improved with 250 ms vs. 50 ms acquisition time (~ 10%). Overall, this analysis indicates that longer MS1 accumulation time provides an advantage to DIA-Umpire algorithm with respect to the total number of identified peptide ions, especially peptide ions identified with a high quality MS1 precursor ion signal.



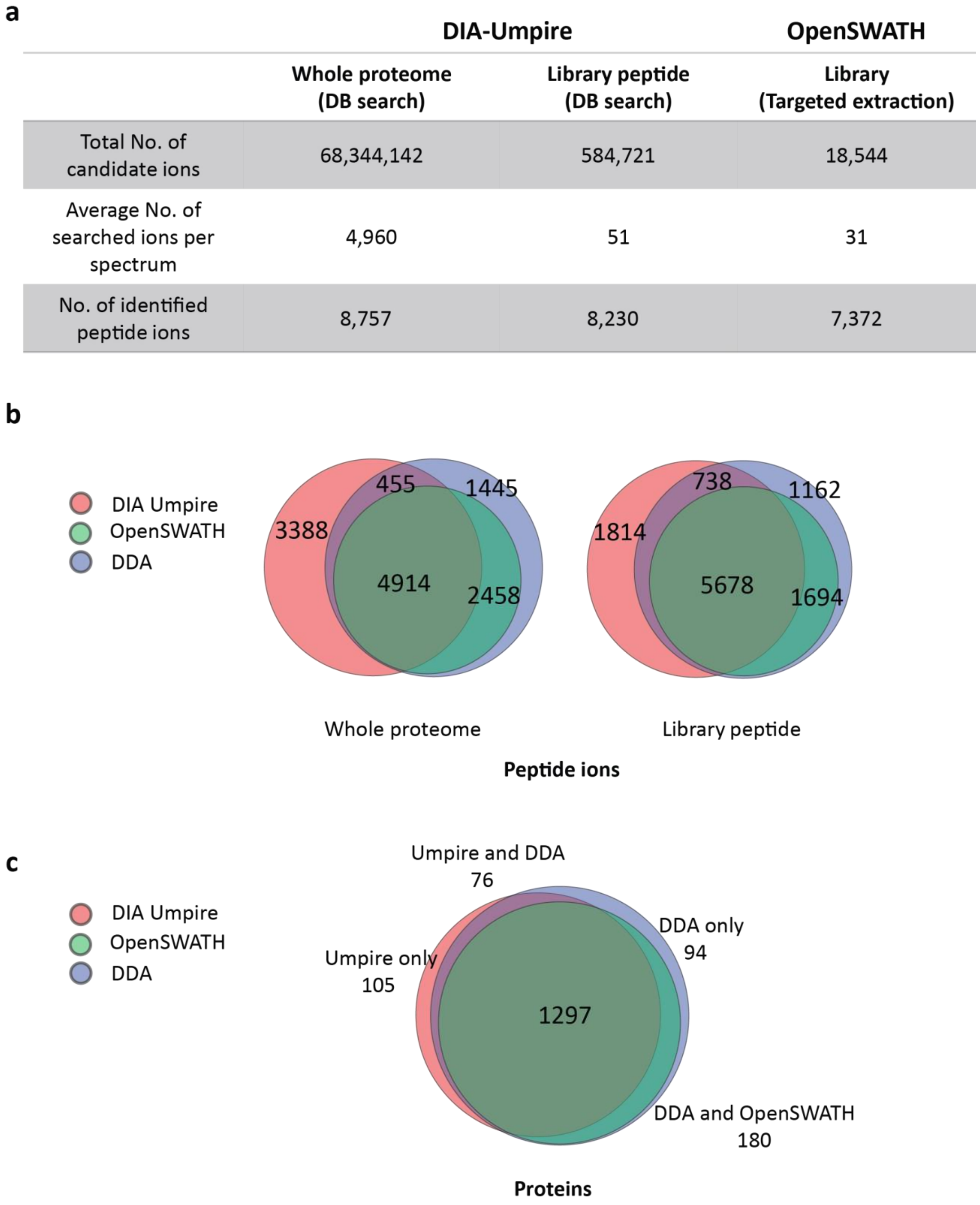
Supplementary Fig. 3. Untargeted peptide identification using DDA and DIA data from human cell lysate samples using three search engines combined. DIA pseudo MS/MS spectra were searched using X! Tandem, Comet, and MSGF+, and combined using iProphet. Protein and peptide ion identifications were then filtered at 1 % FDR using target-decoy approach. **(a)** The numbers of proteins and peptide ions identified at 1% FDR in DDA and in DIA data. *Left*: number of protein identifications in each experiment (1,831 proteins identified from DDA data, 1,692 from DIA, 1,964 in total). *Right*: Total number of peptide ion identifications from two replicates (10,822 peptide ions identified from DDA data, 10,922 from DIA, 14,997 in total). Compared to using X! Tandem only (main text, Figure 4) when the results from all three search engines were combined the number of identifications increased in both DDA (by 17% and 11% for peptide ions and proteins, respectively) and in DIA data (by 25% and 16% for peptide ions and proteins, respectively). However, the overlap between the DIA and DDA identified peptide ions and proteins increased only slightly, to 45% and 79%. Of the peptide ions identified by DIA and not DDA at 1% FDR (total 4,175 peptide ions), the majority of the remaining peptide ions were not identified by DDA because no MS/MS was acquired (2,742). **(b)** Percent of fragments ions matched in pseudo MS/MS spectra extracted from DIA data as a function of the MS1 peptide ion identity in DDA data. Data points (peptide ions) and the summary density plots are labeled according to the three categories of peptide ions: ions identified from DIA data at 1% FDR ("Identified in DIA"; blue), and unidentified in DIA (orange; these ions were located in DIA data as described in Online Methods). **(c)** Comparison between DDA and DIA in terms of numbers of fragments matched among two categories of peptide ions, showing that peptide ions identified with confidence from DDA but not DIA have fewer fragment ions that could be matched.



Supplementary Fig. 4. Untargeted peptide identification using DDA and DIA data from *E. coli* cell lysate samples with X! Tandem search engine. Results for *E. coli* data were similar to those obtained for human cell lysate data (see **Fig. 4**).



Supplementary Fig. 5. Untargeted peptide identification using DDA and DIA data from *E. coli* cell lysate samples with three search engines combined. Results for *E. coli* data were similar to those obtained for human cell lysate when using X! Tandem, Comet, and MSGF+ (combined using iProphet; see **Supplementary Fig. 3**).

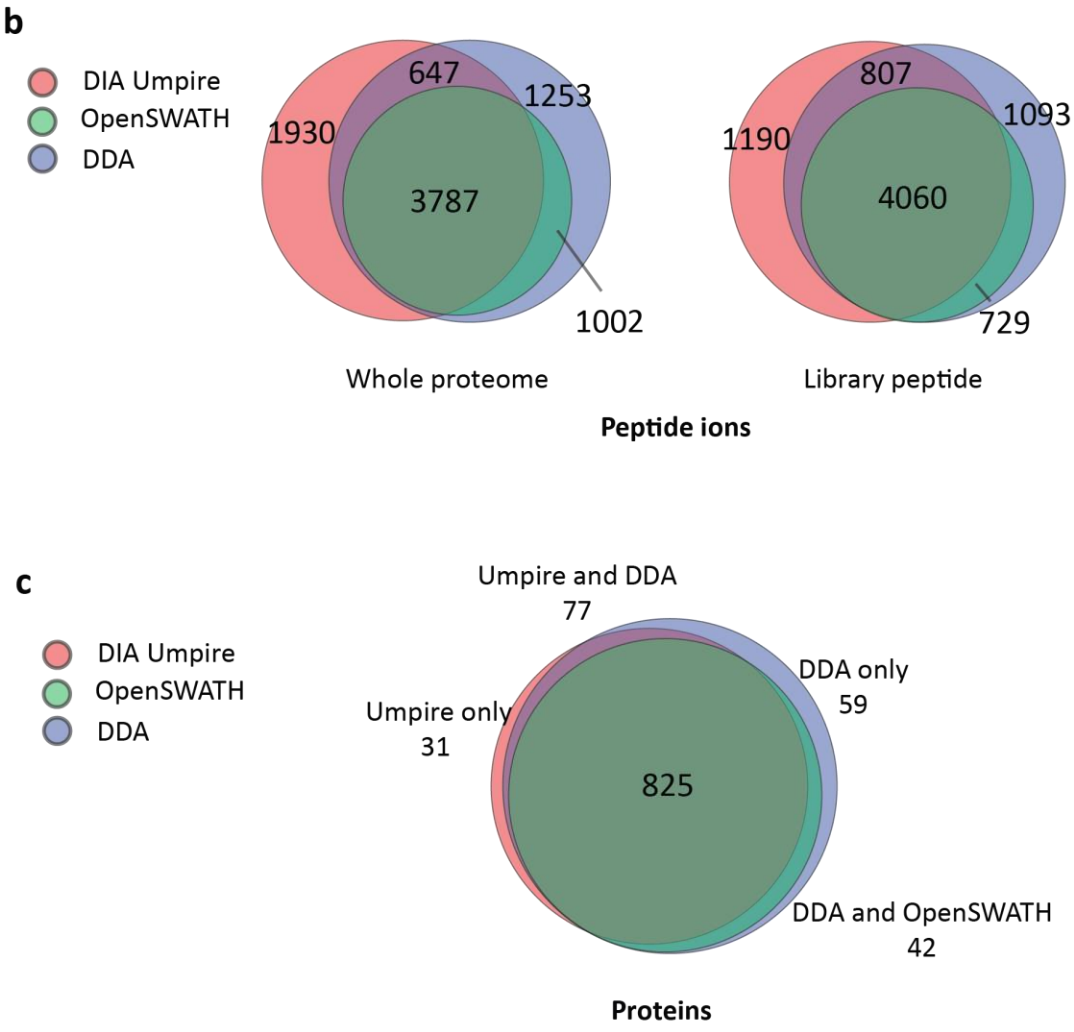


	DIA-Umpire		OpenSWATH
	Whole proteome (DB search)	Library peptide (DB search)	Library (Targeted extraction)
Total No. of candidate ions	68,344,142	584,721	18,544
Average No. of searched ions per spectrum	4,960	51	31
No. of identified peptide ions	8,757	8,230	7,372

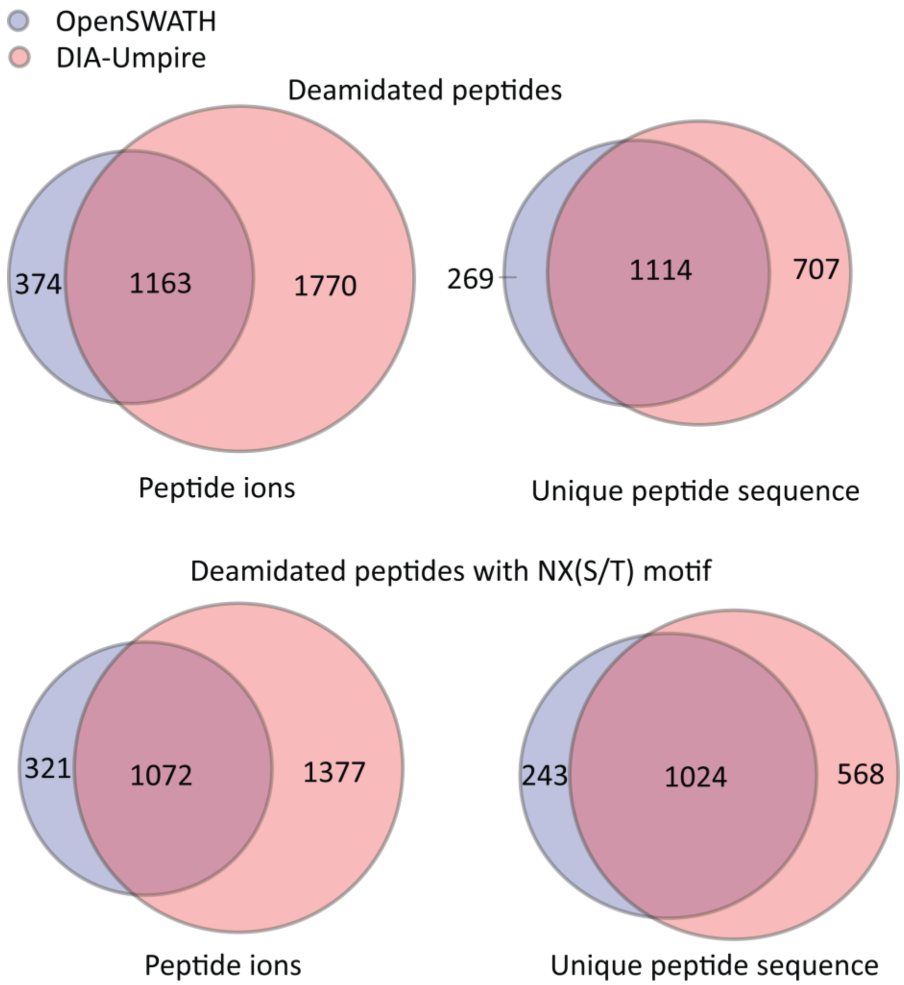
Supplementary Fig. 6 Comparison between untargeted DIA-Umpire analysis and OpenSWATH targeted extraction: effect of the search space. Human cell lysate data (a) The pseudo MS/MS spectra extracted using DIA-Umpire were searched against two sequence databases: “Whole proteome” contains all proteins in the human proteome (plus decoy proteins); “Library peptide” database contains only the sequences of the DDA identified peptides (i.e. it is built using the same peptides as the spectral library used for targeted extraction with OpenSWATH). The total numbers of candidate peptide ions considered for scoring of pseudo MS/MS spectra extracted by DIA-Umpire during database search against “Whole proteome” or “Library peptide” databases were estimated using the following parameters: 30 ppm precursor mass tolerance; peptide sequence length: 4 ~ 50 amino acids; one missed cleavage allowed; charge state considered: 2+, 3+, or 4+; *m/z* range: 350 ~ 1200 Da; variable modifications: oxidation of methionine, cysteine alkylation, conversion of pyroglutamate from glutamine or glutamic acid, and n-terminal acetylation (allowing less than six modifications on the same peptide). For OpenSWATH analysis, the following parameters were used to estimate number of candidate fragment groups in the experimental SWATH MS2 data considered for each target library peptide: mass range of the corresponding SWATH *m/z* isolation (25 Da wide) and ± 1 minute retention time window. The use of the precursor ion *m/z* value from MS1 or MS2 unfragmented precursors as a constraint during database search was the primary factor contributing to the significant reduction in the number of candidate peptide ions considered for scoring against each spectrum (from 68,344,142 peptides in the whole proteome database to 4,960 searched ions per spectrum on average, i.e. 13,779 fold reduction). Because targeted data extraction in OpenSWATH used the retention time of the peptide and wide (25 Da) *m/z* SWATH selection window (but not the precursor peptide *m/z*) for constraining the “search space”, the reduction in the number of candidates was less significant (from 18,544 peptide ions in the library to 31 ions per ± 1 minute retention time slice of the corresponding 25 Da MS2 SWATH scan, i.e. 598 fold reduction). The order of magnitude difference in the search space reduction in DIA-Umpire/’Whole proteome’ analysis (compared to OpenSWATH/Library analysis) explains why DIA-Umpire untargeted analysis performed well. **(b)** Venn diagram of peptide ion identifications among the three analyses. DIA-Umpire/Whole proteome and OpenSWATH/Library identified a comparable number of peptide ions, but the two methods had only a moderate overlap. OpenSWATH identified larger fraction of peptides in the library, 79% (i.e. (4,914 + 2,458 = 7,372) / 9,272) vs. 58% (i.e. (4914 + 455 = 5,369) / 9,272)) for DIA-Umpire. At the same time, DIA-Umpire was able to identify a large number of peptide ions not present in the library. DIA-Umpire/’Library peptide’ analysis had an effective search space similar to that of OpenSWATH, resulting in even closer performance (and better overlap) between the two methods: the overlap between the DIA-Umpire identified peptides and the DDA-identified peptides improved to 69% (or (5,678 + 738) / 9,272 peptide ions) **(c)** Venn diagram of protein identifications among the three analyses (Whole proteome sequence database was used for DIA-Umpire).

a

	DIA-Umpire		OpenSWATH
	Whole proteome (DB search)	Library peptide (DB search)	Library (OpenSWATH)
Total No. of candidate ions	7,026,825	315,465	13,378
Average No. of searched ions per spectrum	500	39	35
No. of identified peptide ions	6,364	6,057	4,789

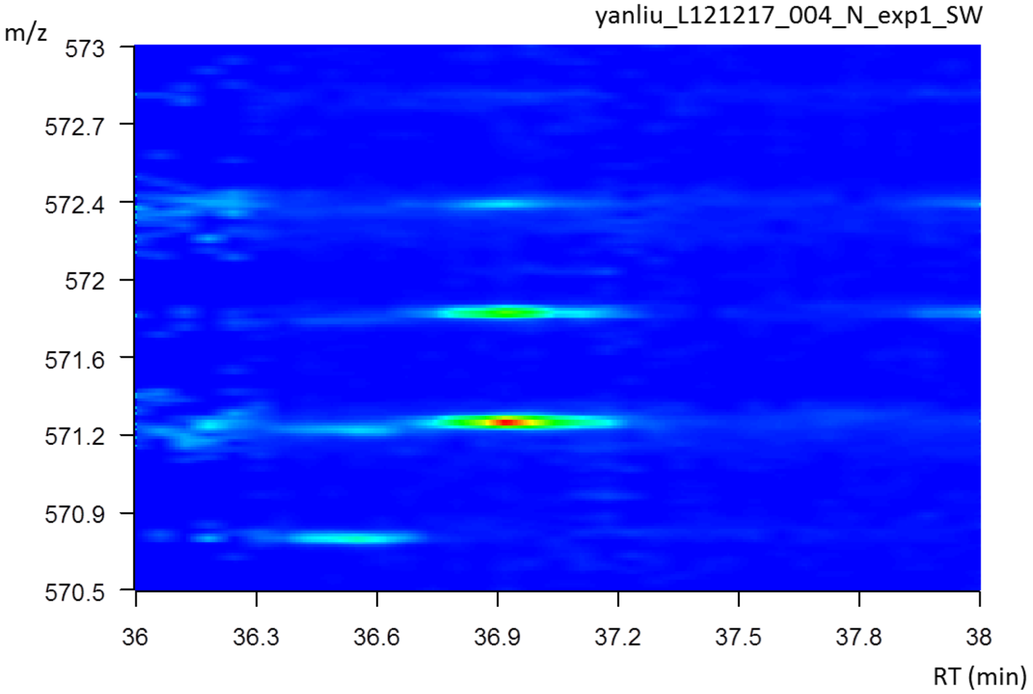


Supplementary Fig. 7 Comparison between untargeted DIA-Umpire analysis and OpenSWATH targeted extraction: effect of the search space. *E. coli* cell lysate data. Results similar to those presented from human cell lysate were obtained for *E. coli* data (see **Supplementary Fig. 8**)



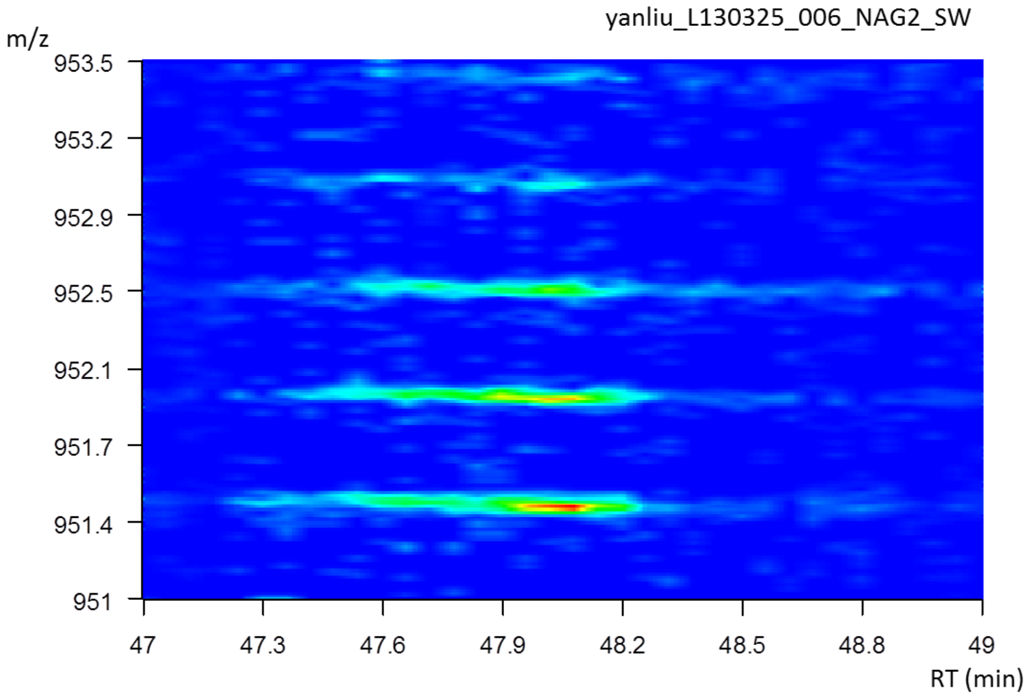
Supplementary Fig. 8 Deamidated peptide identifications (the number of peptide ions and unique peptide sequences) from DIA-Umpire and OpenSWATH targeted search. Glycoproteomics data from Liu *et al. Mol Cell Proteomics* (2014). Upper panel: all peptides. *Lower panel*: peptides containing the NX(S/T) motif expected to be significantly enriched in these data.

Sequence	m/z	Charge	OpenSWATH RT	mProphet m_score	DIA-Umpire RT
NTTFNVESTK	571.76	2	36.9	3.03E-05	N/A
NTTFNVESTK	571.27	2	36.9	1.07E-07	36.88



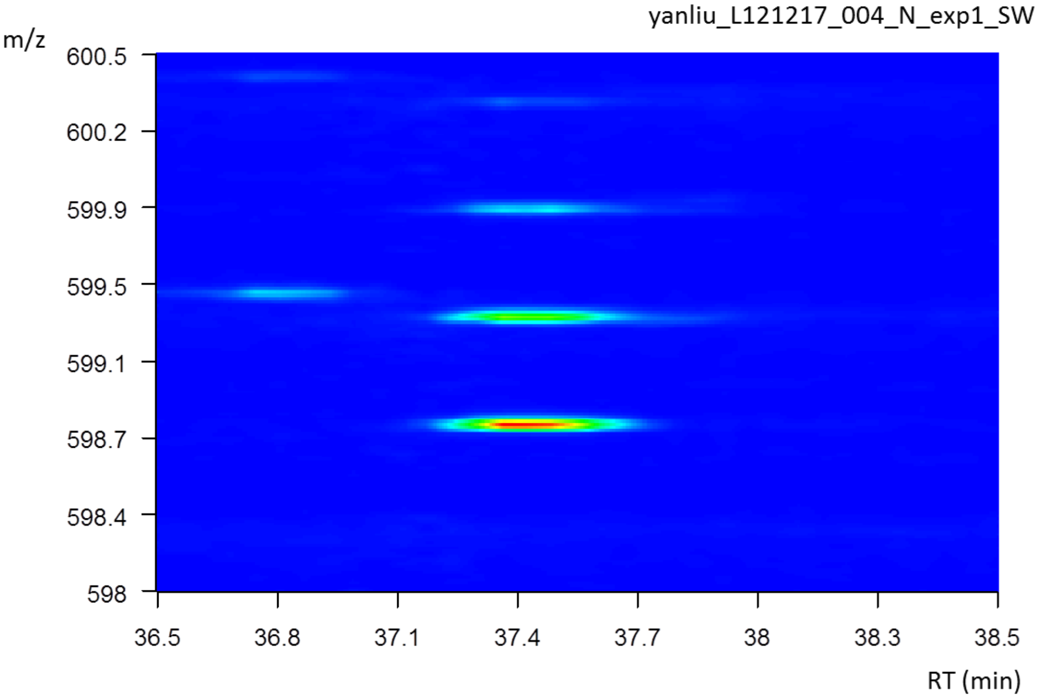
Supplementary Fig. 9 Example of an ambiguous identification involving the deamidated peptide NTTFNVESTK by OpenSWATH targeted search. Although retention times of different modified/unmodified peptide species can help resolve the ambiguity, how different modifications influence retention time remains an open problem for computational prediction. If both species exist in the library and only one of them is present in the sample, library spectra from both species might match the same fragment peak groups in the queried DIA MS2 data. Although the correct one should get a higher matching score, the score of the incorrect one is likely to be better than any decoy peptide and thus also deemed a confident identification. In the glycoproteomics application presented in this work, identification of N-linked glycopeptides depends on detection of asparagine deamidation in peptides due to PNGase F treatment, which causes only a small mass shift (0.984 Da), resulting in both modified and unmodified peptides being co-fragmented. Therefore, we searched the OpenSWATH data for identifications of both deamidated and non-deamidated species reported as highly confident (m_score < 0.01), at the same retention time (within 1 second), and of the same charge state. Here we present one example (see **Supplementary Figs. 10–13** for additional examples) in which OpenSWATH was not able to distinguish deamidated peptide ions from unmodified peptide ions (we manually checked these by using the exact precursor mass). More specifically, two separate identifications with different modification site compositions (with one and two deamidations; modification site shown in red) were reported by OpenSWATH. The two identifications both had an extremely small m_score (from mProphet), i.e. they both were reported as high confidence identifications. The two identifications had identical retention times. The MS1 signal image shown above suggests there is only one peptide eluting at RT = 36.9 minutes (precursor *m/z* of 571.27 Da). DIA-Umpire reported only one (singly deamidated) form, further supported by the presence of NXS/T motif covering the reported site. This example demonstrates that the knowledge of the precursor mass can be very valuable for differentiating between different modification forms of the same peptide sequence in DIA experiments.

Sequence	m/z	Charge	OpenSWATH RT	mProphet m_score	DIA-Umpire RT
NSPLDEENLTQENQDR	951.91	2	47.98	1.30E-06	N/A
NSPLDEENLTQENQDR	951.42	2	47.98	3.68E-08	48.0



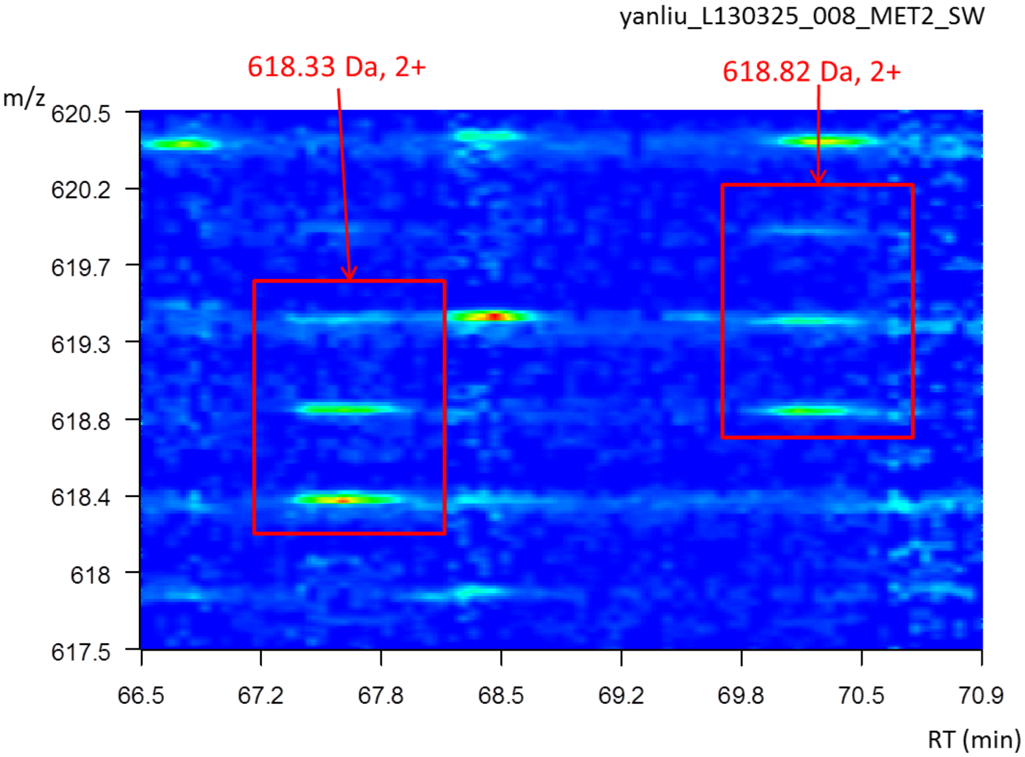
Supplementary Fig. 10 Example of an ambiguous identification of the deamidated peptide NSPLDEENLTQENQDR by OpenSWATH targeted search. Two separate identifications (in unmodified form and in a deamidated form; the site of the modification is shown in red) were reported by OpenSWATH. The two identifications both had an extremely small m_score (from mProphet), i.e. they both were reported as high confidence identifications. The two identifications had identical retention times. The MS1 signal image shown above suggests there is only one peptide eluting at RT = 47.98 minutes (precursor *m/z* of 951.42 Da). DIA-Umpire reported only one (unmodified) form.

Sequence	m/z	Charge	OpenSWATH RT	mProphet m_score	DIA-Umpire RT
DIENFNSTQK	599.26	2	37.41	0.000105616	N/A
DIENFNSTQK	598.77	2	37.39	1.67E-07	37.35



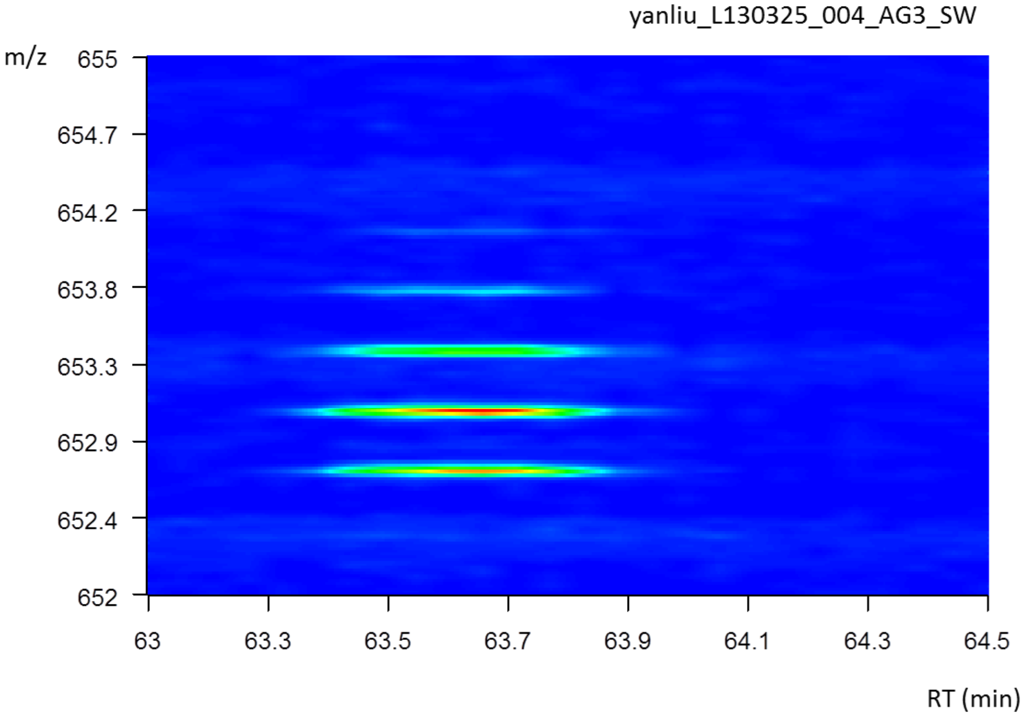
Supplementary Fig. 11 Example of an ambiguous identification of the peptide DIENFNSTQK by OpenSWATH targeted search. Two separate identifications with different modification site compositions (with one and two deamidations; modification site shown in red) were reported by OpenSWATH. The two identifications both had a small m_score (from mProphet), i.e. they both were reported as high confidence identifications. The two identifications had almost identical retention times (within 0.02 minute). The MS1 signal image shown above suggests there is only one peptide eluting at RT = 37.4 minutes (precursor *m/z* of 598.77 Da). DIA-Umpire reported only one (singly deamidated) form, further supported by the presence of NXS/T motif covering the reported site.

Sequence	m/z	Charge	OpenSWATH RT	mProphet m_score	DIA-Umpire RT
TGNGLFLSEGLK	618.82	2	67.58	3.04E-09	69.99
TGNGLFLSEGLK	618.33	2	67.58	3.42E-08	67.59

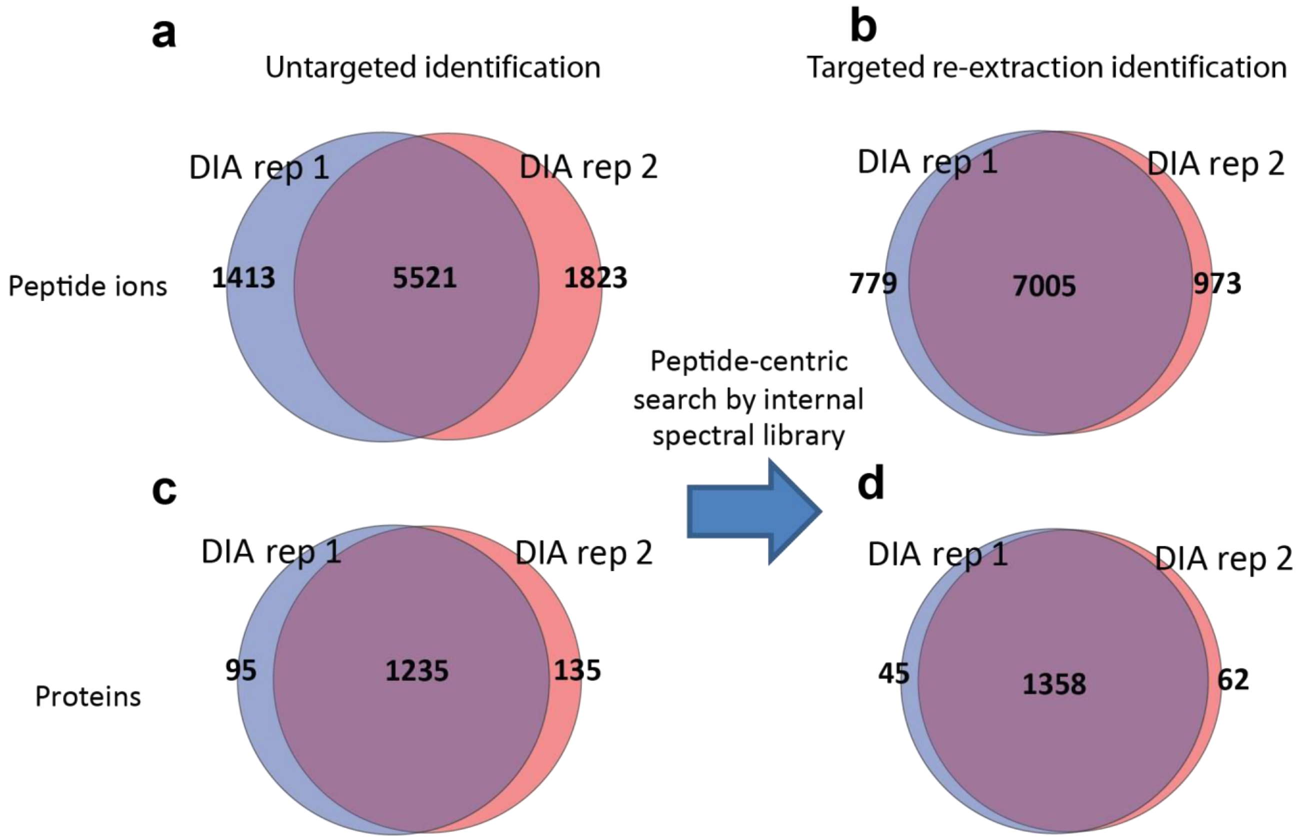


Supplementary Fig. 12 Example of an ambiguous identification involving the deamidated peptide TGNGLFLSEGLK. Two separate identifications were reported (in unmodified and in deamidated form) by OpenSWATH. The two identifications both had an extremely small m_score (from mProphet), i.e. they both were reported as high confidence identifications. The two identifications had identical retention times. The MS1 signal image shown above suggests there is only one peptide eluting at RT = 67.6 minutes (precursor *m/z* of 618.33 Da), which was identified by DIA-Umpire as unmodified peptide. In addition, DIA-Umpire identified the deamidated form of the peptide at retention time of 69.99 minutes (also marked on the MS1 signal image).

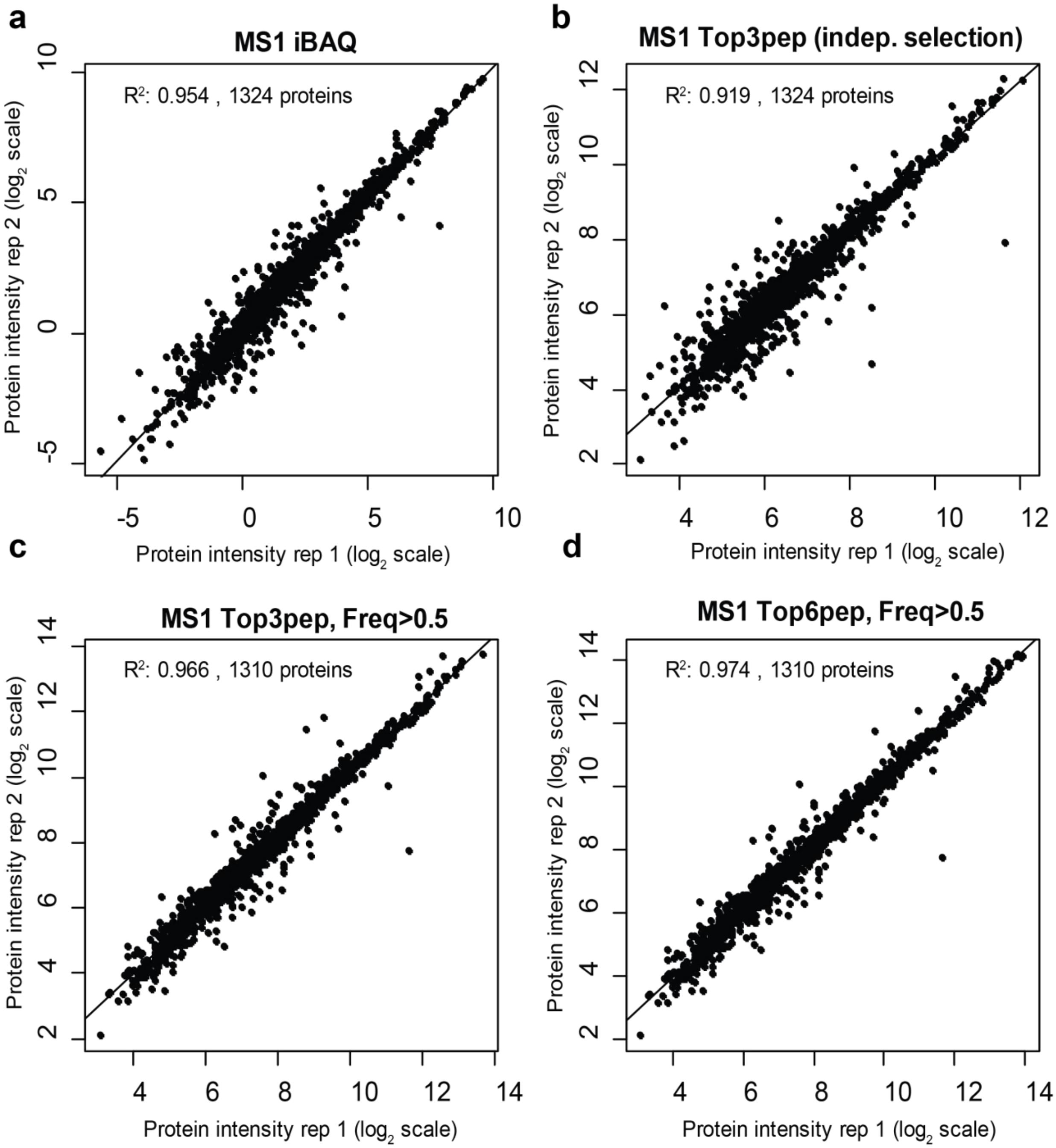
Sequence	m/z	Charge	OpenSWATH RT	mProphet m_score	DIA-Umpire RT
VAPEEHPTLLTEAPLNPK	653.01	3	63.586	0.001410359	N/A
VAPEEHPTLLTEAPLNPK	652.686	3	63.591	3.67E-08	63.58



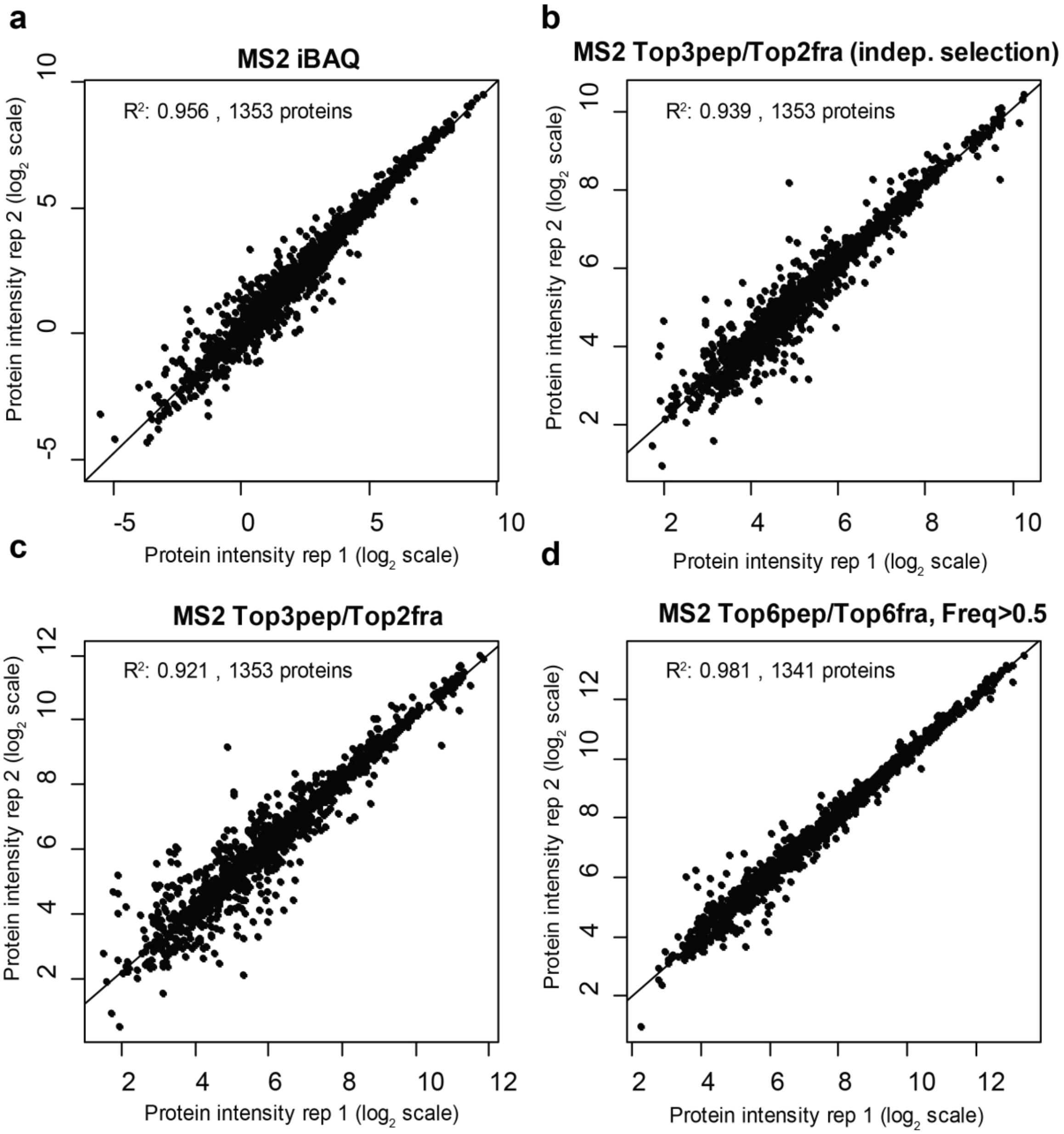
Supplementary Fig. 13 Example of an ambiguous identification of the deamidated peptide VAPEEHPTLLTEAPLNPK by OpenSWATH targeted search. Two separate identifications (in unmodified form and in a deamidated form; the site of the modification is shown in red) were reported by OpenSWATH. The two identifications both had an extremely small m_score (from mProphet), i.e. they both were reported as high confidence identifications. The two identifications had almost identical retention times. The MS1 signal image shown above suggests there is only one peptide eluting at RT = 63.58 minutes (precursor *m/z* of 652.68 Da). DIA-Umpire reported only one (unmodified) form.



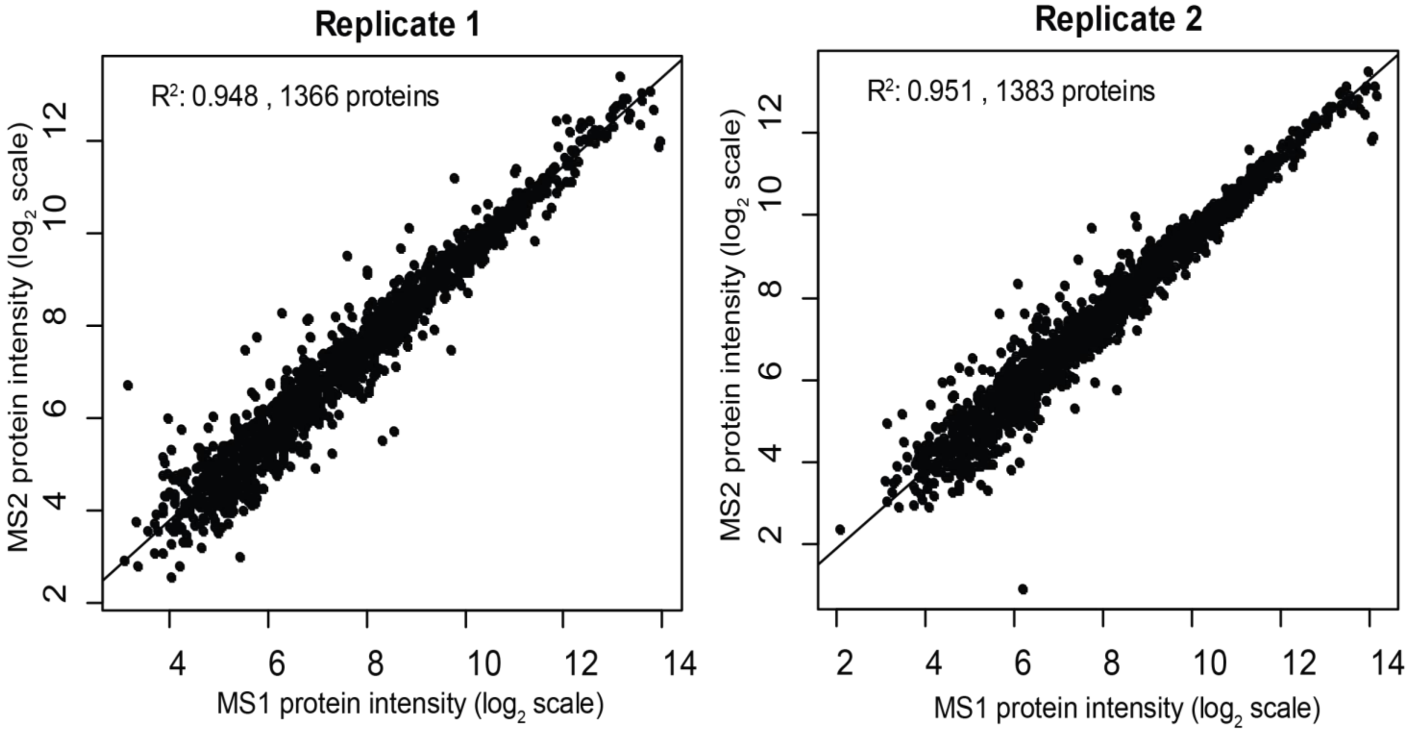
Supplementary Fig. 14 Increased identification coverage after targeted re-extraction in DIA-Umpire. Human cell lysate DIA data**. (a)** Venn diagram of peptide ion identifications in two replicates of human cell lysate DIA (SWATH) data from untargeted X! Tandem search; **(b)** the number of peptide ions identified in both replicates increased after targeted re-extraction; **(c)** same as (a) at the protein level; **(d)** same as (b) at the protein level.



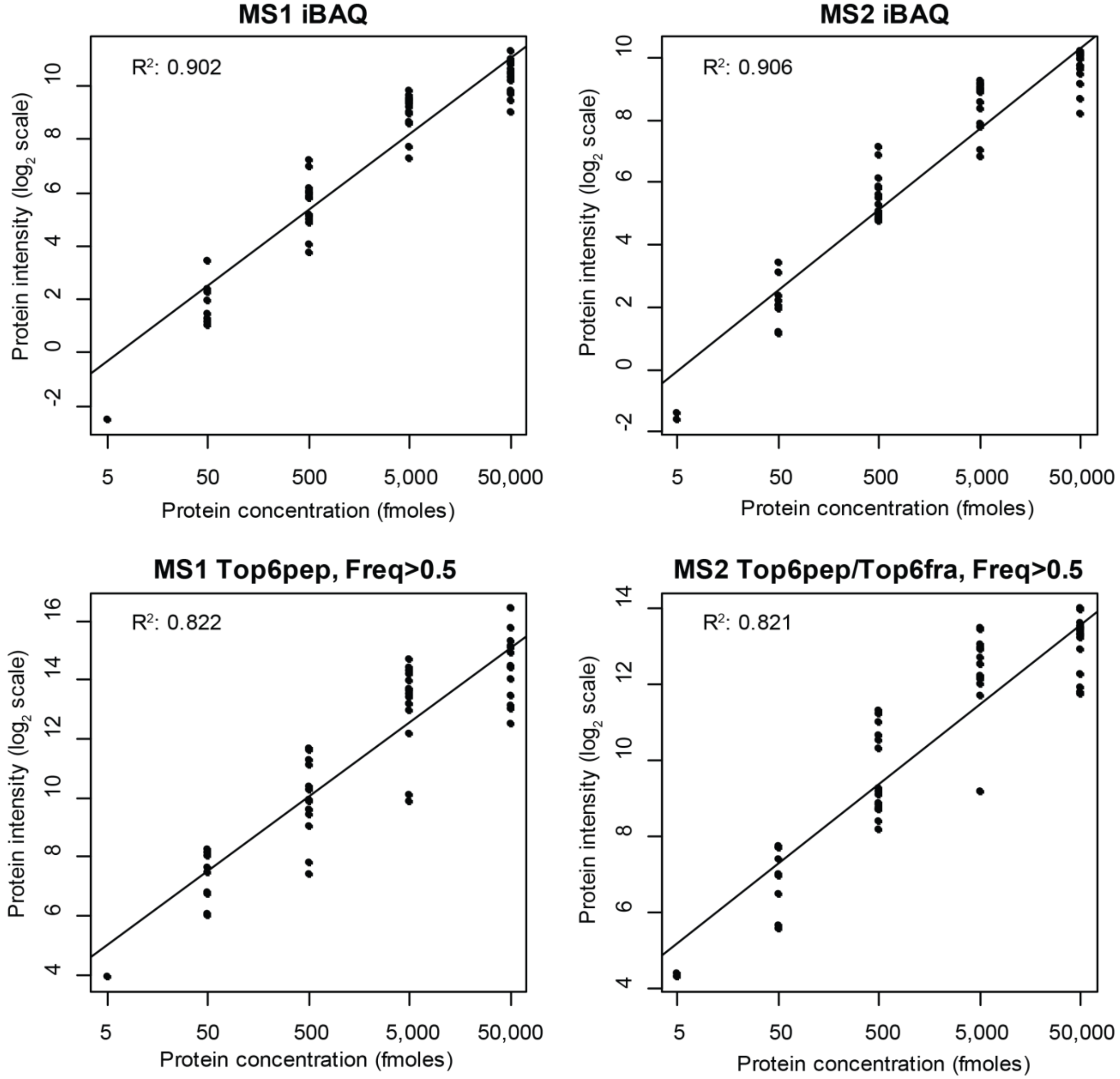
Supplementary Fig. 15 MS1-based protein quantification in DIA human cell lysate data. (a) "MS1 iBAQ" intensity; **(b)** "MS1 Top3pep (indep. selection)": protein intensity estimated by summing the top three most intense peptide ions, independently in each DIA run; **(c)** "MS1 Top3pep, Freq>0.5": same as (b), with an additional requirement that the selected peptides are identified in more than 50% of the runs (Freq > 0.5) in which the corresponding protein was identified; **(d)** "MS1 Top6pep, Freq>0.5": same as (c), but using six most intense peptides. Note that selection of consistently identified peptide ions (Freq > 0.5 filter) significantly improves the reproducibility of protein intensities between the replicates.



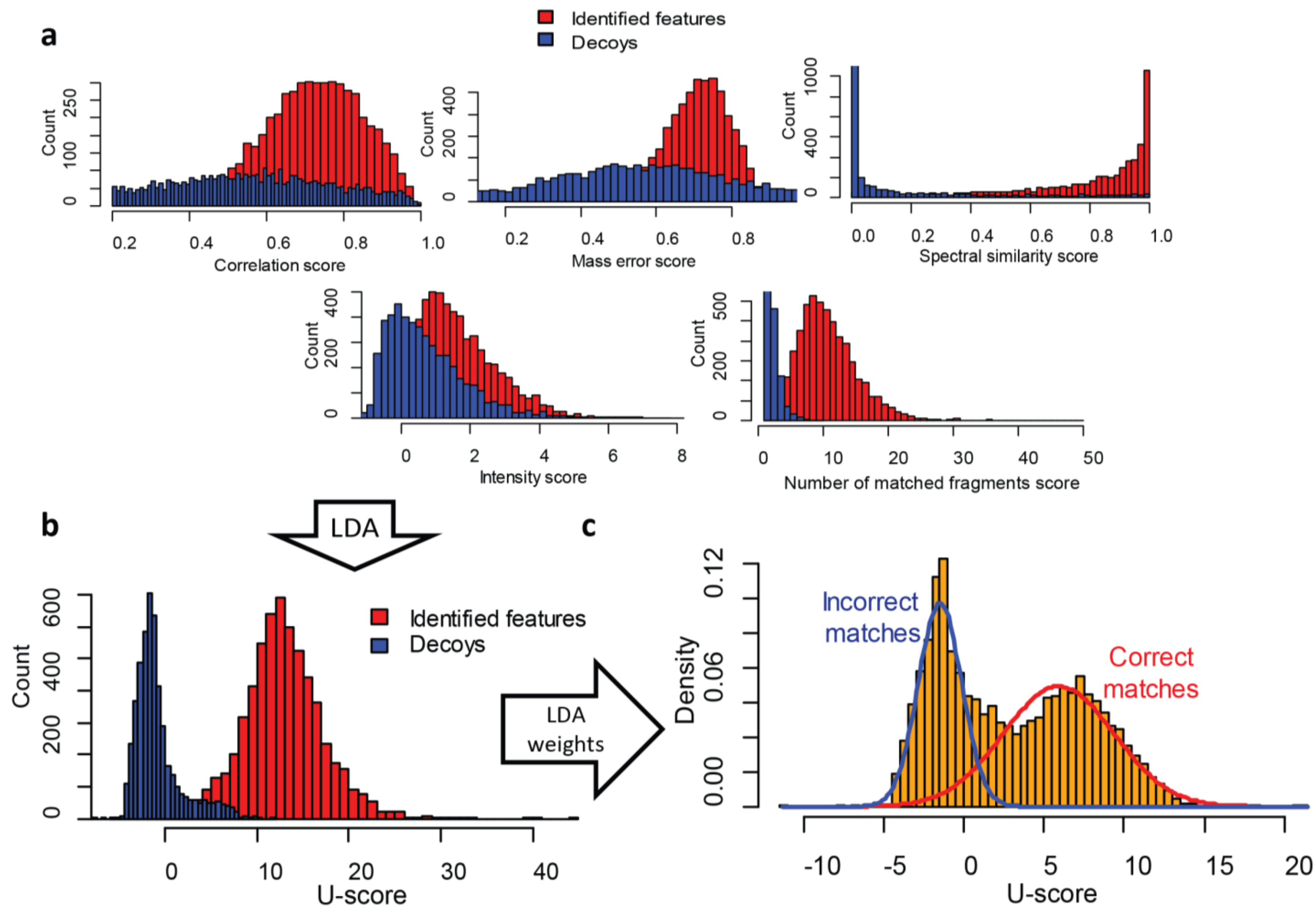
Supplementary Fig. 16 MS2-based protein quantification in DIA human cell lysate data. (a) "MS2 iBAQ" intensity; **(b)** "MS2 Top3pep/Top2fra (indep. selection)": protein intensity estimated by summing the top three most intense peptide ions. Peptide ion intensities are estimated by summing the intensities of their top 2 most intense matched fragments. Protein intensities are computed independently for each DIA run; **(c)** "MS2 Top3pep/Top2fra": same as (b), but using peptide ions and fragments having the highest overall intensity across all runs (here, highest summed intensity across the two replicates); **(d)** same as (c), but using six selected peptide ions and fragments, with an additional requirement that the selected peptides and fragments are identified in more than 50% of the runs (Freq > 0.5) in which the corresponding protein (peptide selection step) or peptide (fragment selection step) was identified. Note that selection of consistently identified peptide ions and fragments (Freq > 0.5 filter) significantly improves the reproducibility of protein intensities between the replicates.



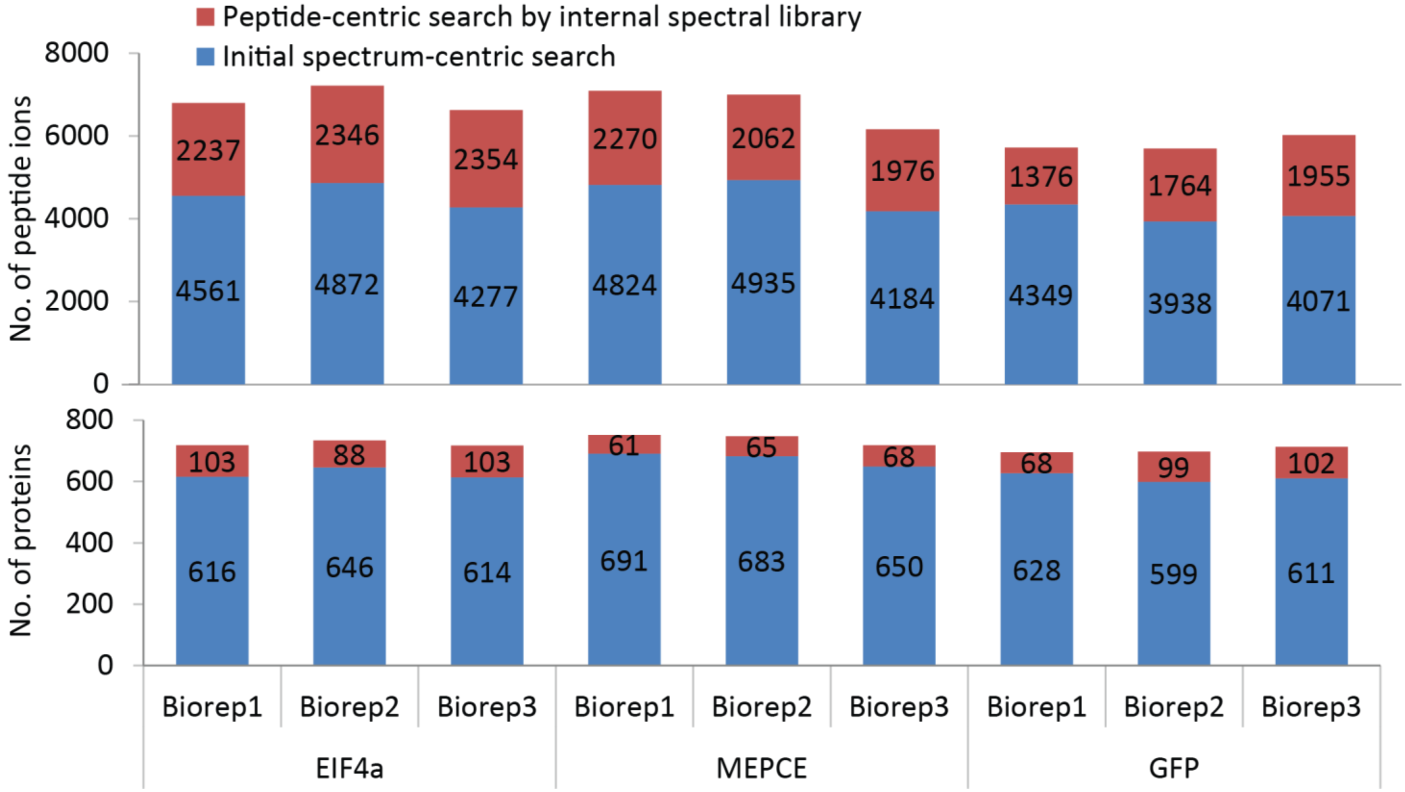
Supplementary Fig. 17 Comparison between MS1 and MS2-based protein quantification. Human cell lysate data. MS1-based intensities are computed using the 'Top6pep, Freq>0.5' method. MS2-based protein intensities are computed using the "Top6pep/Top6fra, Freq>0.5" method.



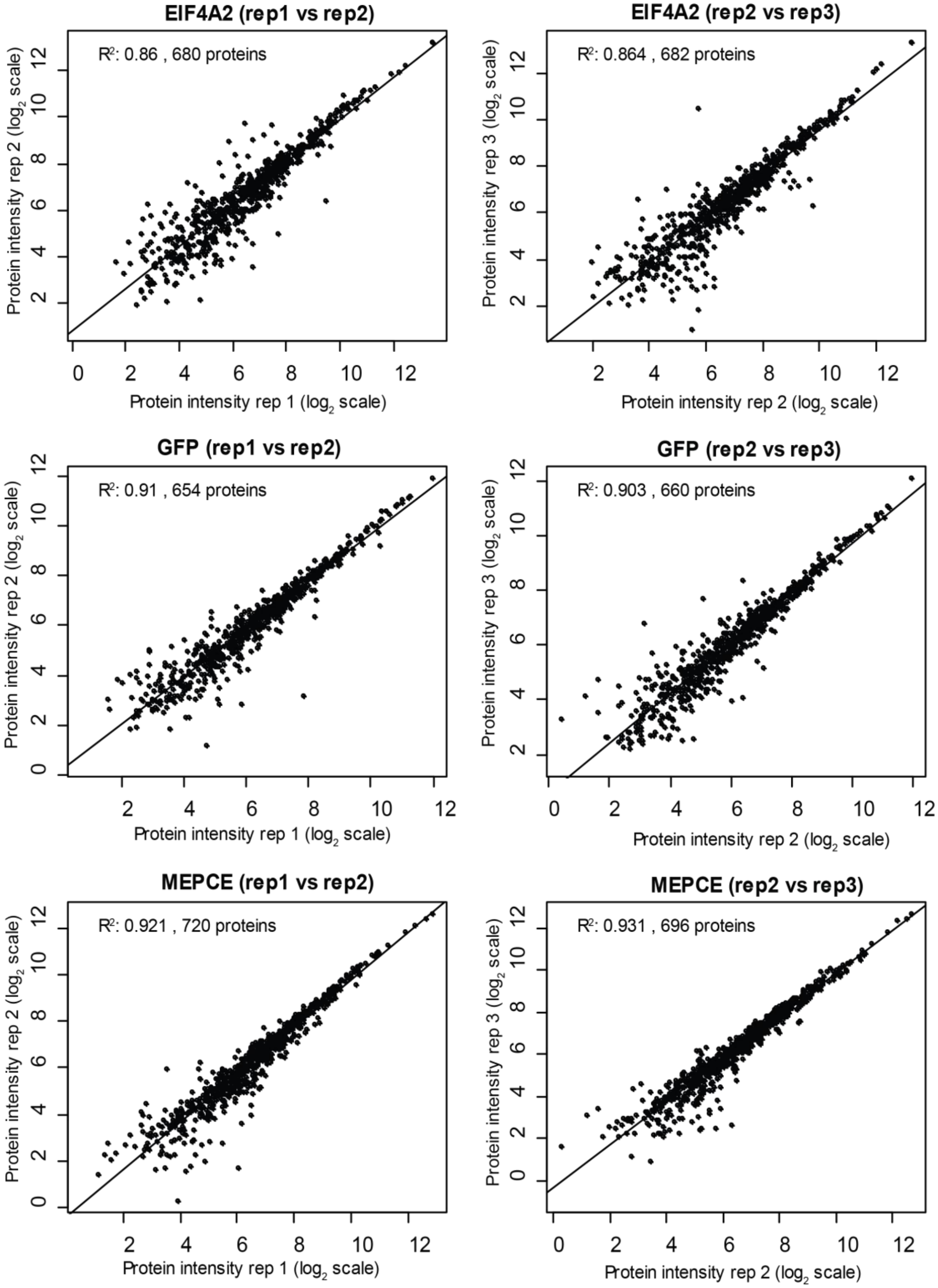
Supplementary Fig. 18 Protein quantification results in the UPS2 standard protein sample. In UPS2 samples, protein concentrations are known and span five orders of magnitude (8 proteins in each of the five abundance bins; no proteins were quantified in the lowest abundance bin in this study). Proteins were quantified using four quantification methods. **(a)**"MS1 iBAQ" intensity; **(b)**"MS2 iBAQ" intensity; **(c)**"MS1 Top6pep, Freq>0.5": protein intensity estimated by summing the top six most intense peptide ions which are consistently identified (Freq > 0.5 filter); **(d)**"MS2 Top6pep/Top6fra, Freq>0.5" : protein intensity estimated by summing the top six most intense peptide ions with an additional requirement that the selected peptides and fragments are identified in more than 50% of the runs (Freq > 0.5) in which the corresponding protein (peptide selection step) or peptide (fragment selection step) was identified.



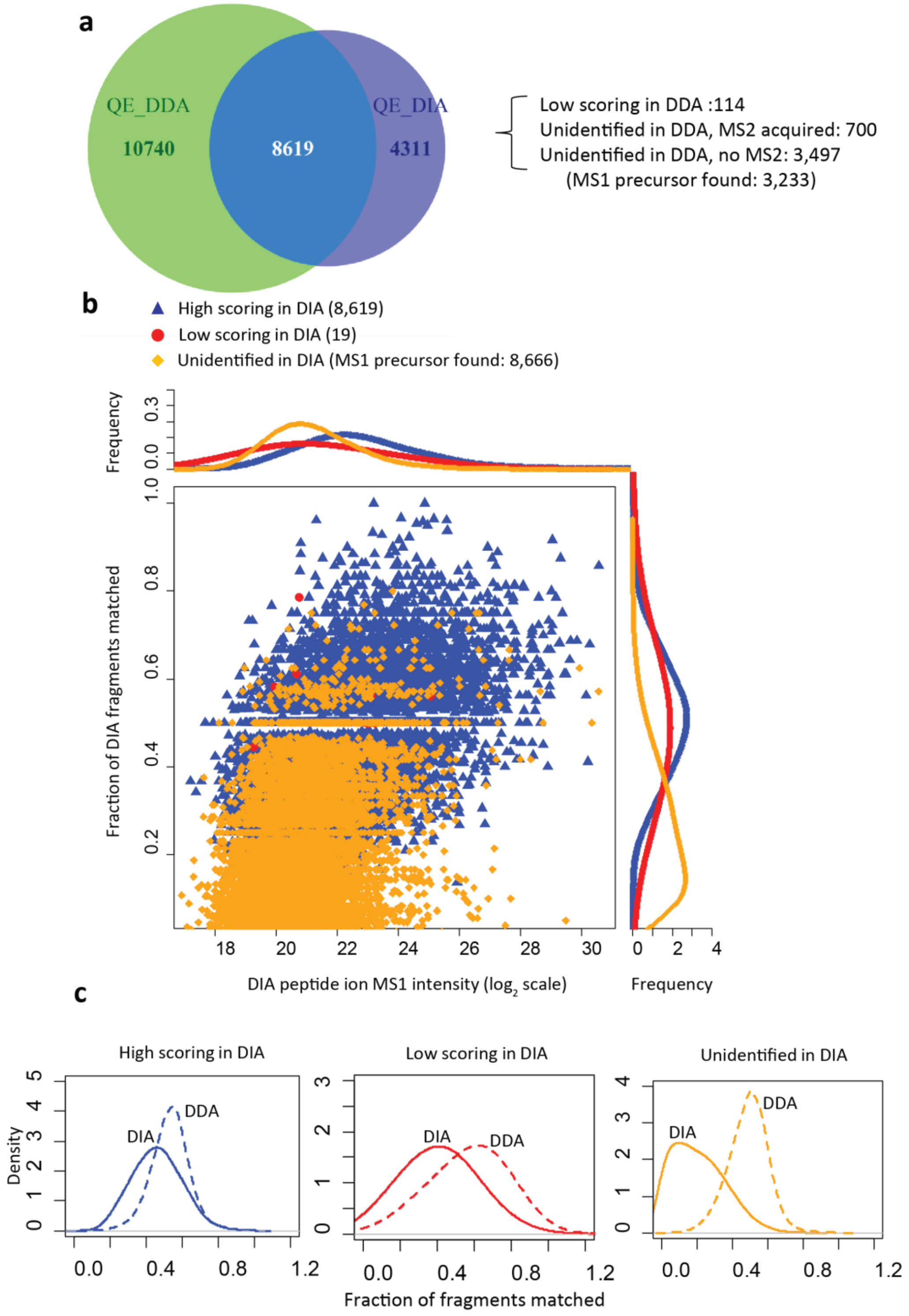
Supplementary Fig. 19 Distributions of scores computed by the targeted re-extraction algorithm of DIA-Umpire. AP-SWATH dataset, MEPCE bait (biological replicate 3). **(a)** Distributions of the five sub-scores for peptide ions from the positive set ('Identified features', i.e. precursor-fragment group to spectrum matches that were confidently identified by the untargeted spectrum-centric search) and from the negative set (precursor-fragment groups matching to decoy spectra); **(b)** Linear discriminant analysis is used to train the weights in the linear combination used to combine the individual sub-scores to compute a single discriminant score (U-score). Shown are the resulting U-score distributions for positive and negative matches in the training set. **(c)** The final distribution of U-scores for all non-decoy matches. The observed distribution is model as a mixture of two underlying distribution representing high scoring, correct matches (red curve) and low scoring, incorrect matches (blue curve). The parameters of the distributions are learned using the expectation maximization mixture modeling algorithm. The posterior probability of a correct match is computed for a given non-decoy spectrum to precursor-fragments group match from the ratio of learned distributions among correct and incorrect matches. By default, peptide ions with a computed probability above 0.99 are considered confidently identified and contribute, together with the peptide ions identified at the initial untargeted identification stage, to protein quantification for their corresponding protein.



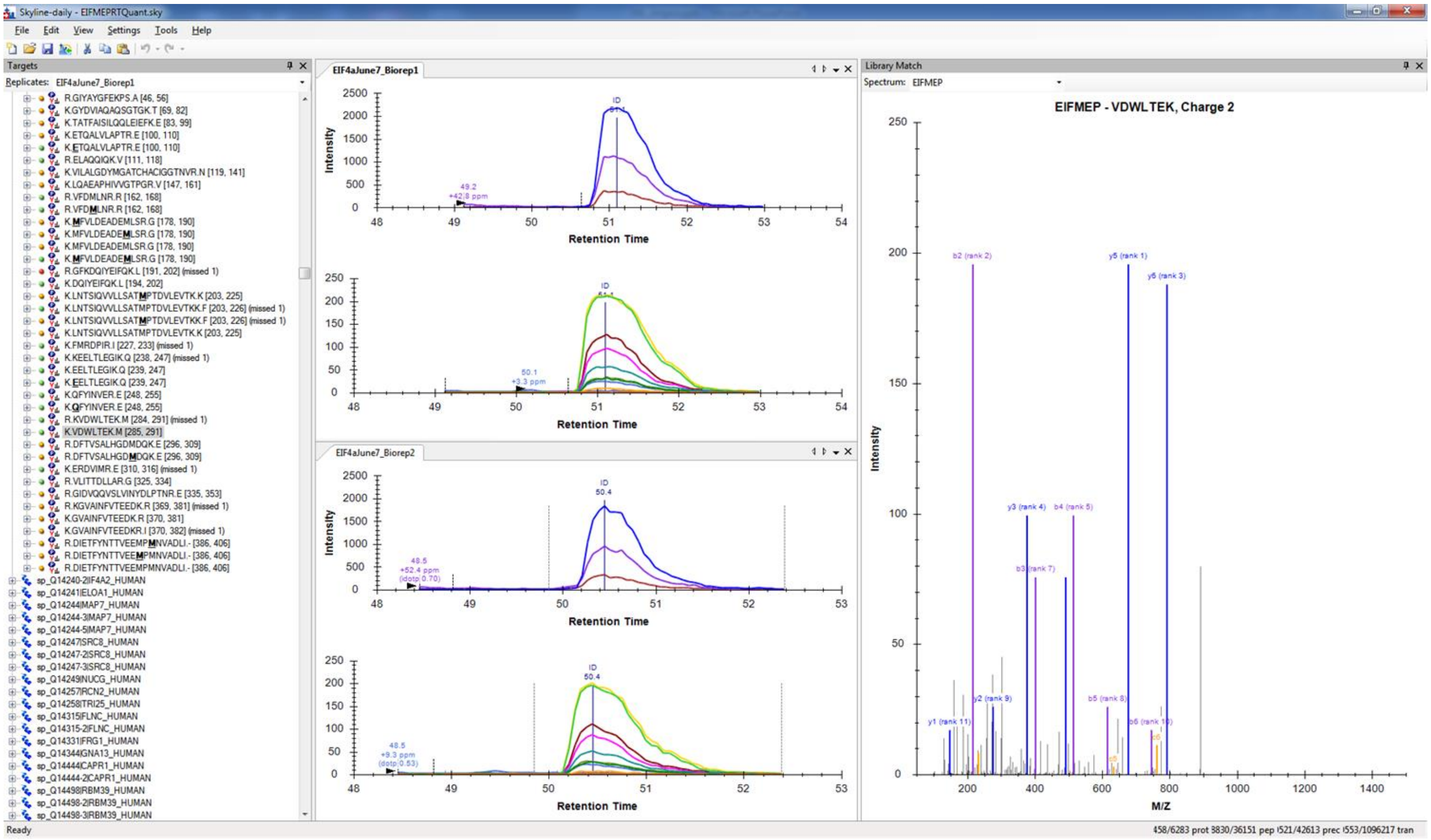
Supplementary Fig. 20 The numbers of identified proteins and peptide ions via untargeted spectrum-centric search and targeted re-extraction matching. The numbers of identified peptide ions and proteins from the initial untargeted (spectrum-centric search) analysis using DIA-Umpire (blue bars) shown separately for 3 biological replicates (Biorep1…Biorep3) of the two bait proteins (EIF4A2 and MEPCE) and the GFP negative controls. Also shown (red bars) the numbers of additional peptide ions and proteins identified by the targeted re-extraction using the spectral library internally generated from the initial search results.



Supplementary Fig. 21 Protein quantification in AP-SWATH data. Protein intensities are computed using the “MS2 Top6pep/Top6fra, Freq>0.5’ approach. Each dot represents computed protein intensities for the same protein in two different biological replicates for the same bait (or GFP control).



Supplementary Fig. 22 Comparison between DDA and DIA for human cell lysate data generated on a Q Exactive Plus instrument. The samples were prepared in the same way as the samples that were analyzed using the AB SCIEX 5600 instrument used to generate the main datasets used in the manuscript. Peptide samples were analyzed using an EasySpray column (25cm x 75um x 2um C18) with 90 minute gradient at 300nl/min coupled online to a Q Exactive Plus (QE) instrument. **(a)** Number of identified peptide ions between QE DDA and QE DIA. **(b)** Detailed analysis for fragment loss and MS1 intensity for different categories of peptide ions: high scoring DIA (blue), low scoring DIA (red), and unidentified in DIA (orange). **(c)** A detailed comparison between DDA and DIA in terms of the numbers of fragments matched among three categories of peptide ions.



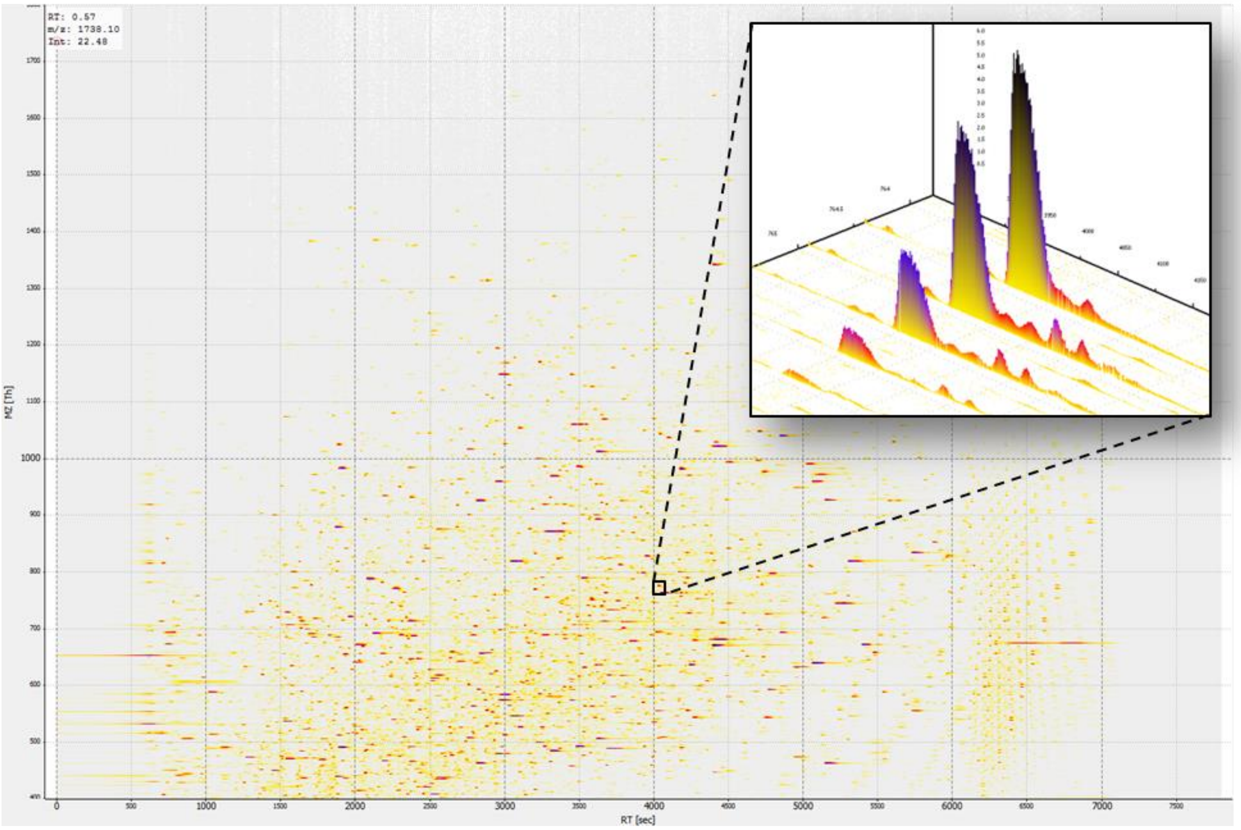
Replicates: EIF4aJune7_Biorep1

R.GIYAYGFEKPS.A [46, 56]
K.GYDVIAQAQSGTGK.T [69, 82]
K.TATFAISILQQLEIEFK.E [83, 99]
K.ETQALVLAPTR.E [100, 110]
K.**E**TQALVLAPTR.E [100, 110]
R.ELAQQIQK.V [111, 118]
K.VILALGDYMGATCHACIGGTNVR.N [119, 141]
K.LQAEAPHIVVGTPGR.V [147, 161]
R.VFDMLNR.R [162, 168]
R.VFD**M**LNR.R [162, 168]
K.**M**FVLDEADEMLSR.G [178, 190]
K.MFVLDEADE**M**LSR.G [178, 190]
K.MFVLDEADEMLSR.G [178, 190]
K.**M**FVLDEADE**M**LSR.G [178, 190]
R.GFKDQIYEIFQK.L [191, 202] (missed 1)
K.DQIYEIFQK.L [194, 202]
K.LNTSIQVVLLSAT**M**PTDVLEVTK.K [203, 225]
K.LNTSIQVVLLSATMPTDVLEVTKK.F [203, 226] (missed 1)
K.LNTSIQVVLLSAT**M**PTDVLEVTKK.F [203, 226] (missed 1)
K.LNTSIQVVLLSATMPTDVLEVTK.K [203, 225]
K.FMRDPIR.I [227, 233] (missed 1)
K.KEELTLEGIK.Q [238, 247] (missed 1)
K.EELTLEGIK.Q [239, 247]
K.**E**ELTLEGIK.Q [239, 247]
K.QFYINVER.E [248, 255]
K.**Q**FYINVER.E [248, 255]
R.KVDWLTEK.M [284, 291] (missed 1)
K.VDWLTEK.M [285, 291]
R.DFTVSALHGDMDQK.E [296, 309]
R.DFTVSALHGD**M**DQK.E [296, 309]
K.ERDVIMR.E [310, 316] (missed 1)
R.VLITTDLLAR.G [325, 334]
R.GIDVQQVSLVINYDLPTNR.E [335, 353]
R.KGVAINFVTEEDK.R [369, 381] (missed 1)
K.GVAINFVTEEDK.R [370, 381]
K.GVAINFVTEEDKR.I [370, 382] (missed 1)
R.DIETFYNTTVEEMP**M**NVADLI.- [386, 406]
R.DIETFYNTTVEE**M**PMNVADLI.- [386, 406]
R.DIETFYNTTVEEMPMNVADLI.- [386, 406]
sp_Q14240-2|IF4A2_HUMAN
sp_Q14241|ELOA1_HUMAN
sp_Q14244|MAP7_HUMAN
sp_Q14244-3|MAP7_HUMAN
sp_Q14244-5|MAP7_HUMAN
sp_Q14247|SRC8_HUMAN
sp_Q14247-2|SRC8_HUMAN
sp_Q14247-3|SRC8_HUMAN
sp_Q14249|NUCG_HUMAN
sp_Q14257|RCN2_HUMAN
sp_Q14258|TRI25_HUMAN
sp_Q14315|FLNC_HUMAN
sp_Q14315-2|FLNC_HUMAN
sp_Q14331|FRG1_HUMAN
sp_Q14344|GNA13_HUMAN
sp_Q14444|CAPR1_HUMAN
sp_Q14444-2|CAPR1_HUMAN
sp_Q14498|RBM39_HUMAN
sp_Q14498-2|RBM39_HUMAN
sp_Q14498-3|RBM39_HUMAN

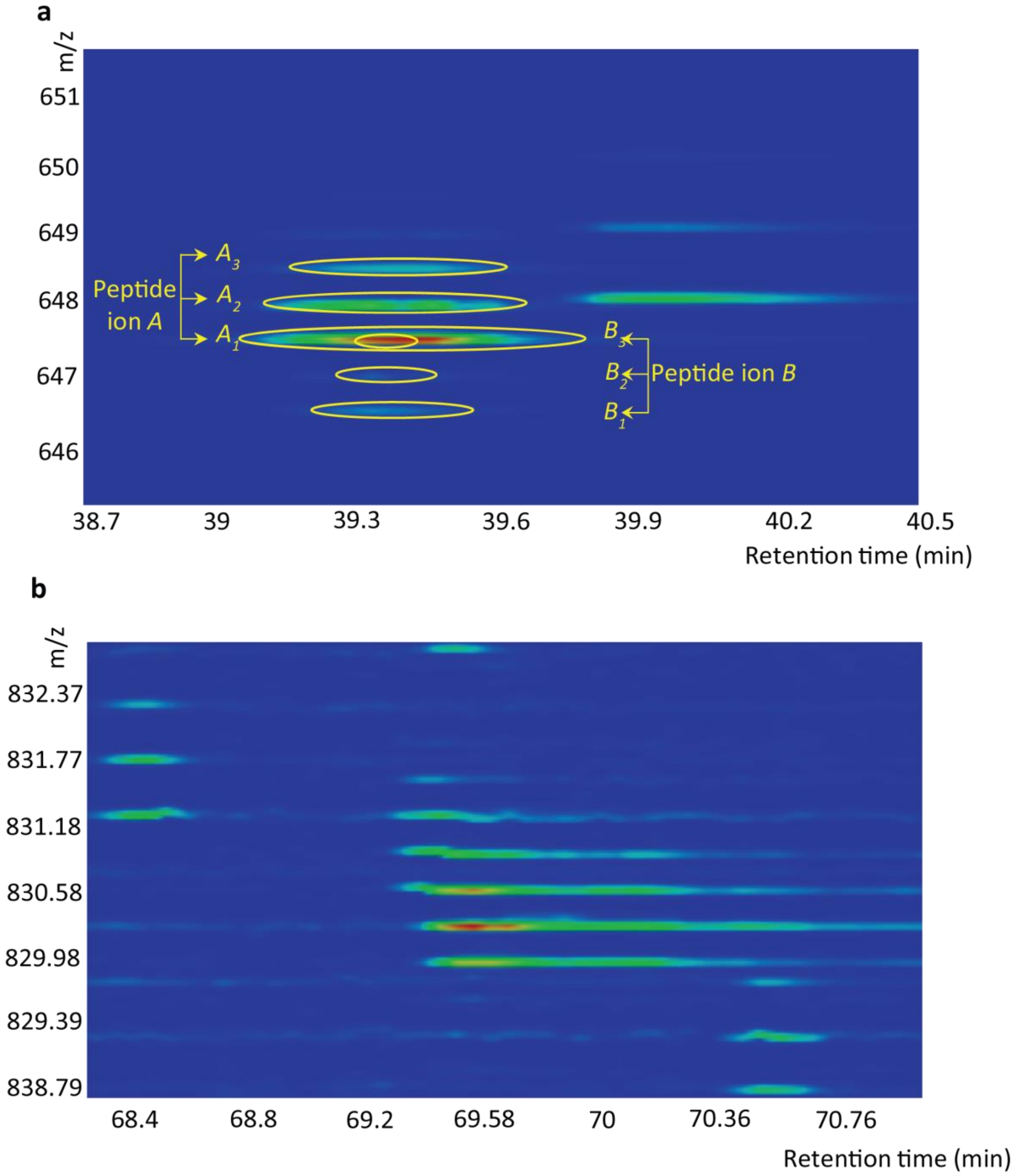
Spectrum: EIFMEP

458/6283 prot 3830/36151 pep 1521/42613 prec 1553/1096217 tran

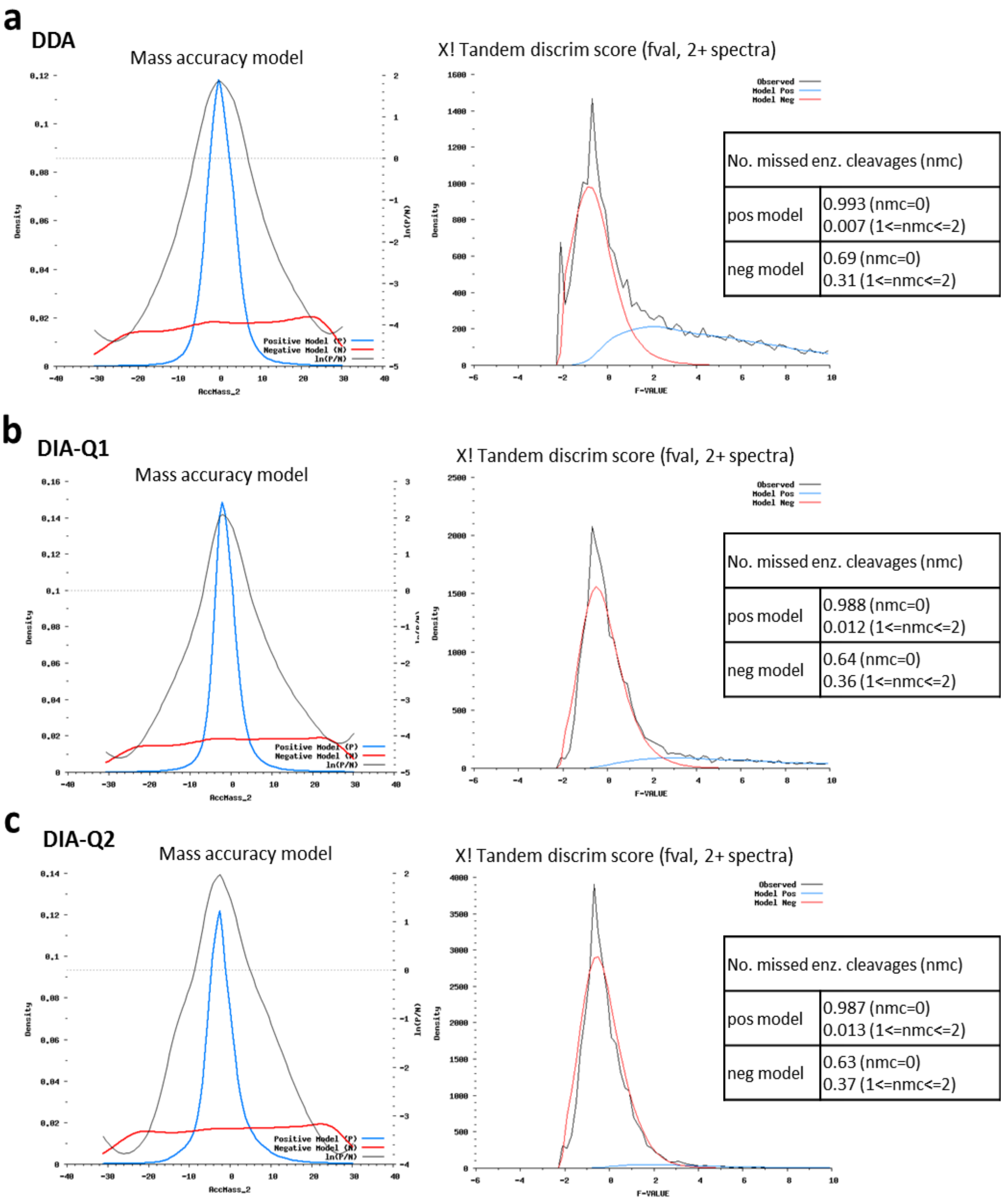
Supplementary Fig. 23 Visualization of DIA signals using DIA-Umpire result files with Skyline.



Supplementary Fig. 24 Illustration of LC-MS data and isotope peaks envelope of a precursor feature. Images exported using OpenMS 1.10.



Supplementary Fig. 25 Examples of co-eluting peptide ions. (a) Two co-eluted peptide ions *A* and *B* with the monoisotopic peak A_1 of peptide ion *A* overlapped with the third isotope peak B_3 of peptide ion B. The peak detection algorithms have a difficulty with detecting B_3 because it is completely buried by A_1 signal. **(b)** Another, more complicated example where co-elution of multiple peptide ions presents an ambiguity with the interpretation of different isotope peak groups. To effectively detect as many true precursor ions as possible, the signal detection algorithm of DIA-Umpire considers each peak curve as a possible monoisotopic peak, and then attempts to find higher isotope peak curves for the assumed monoisotopic peak.

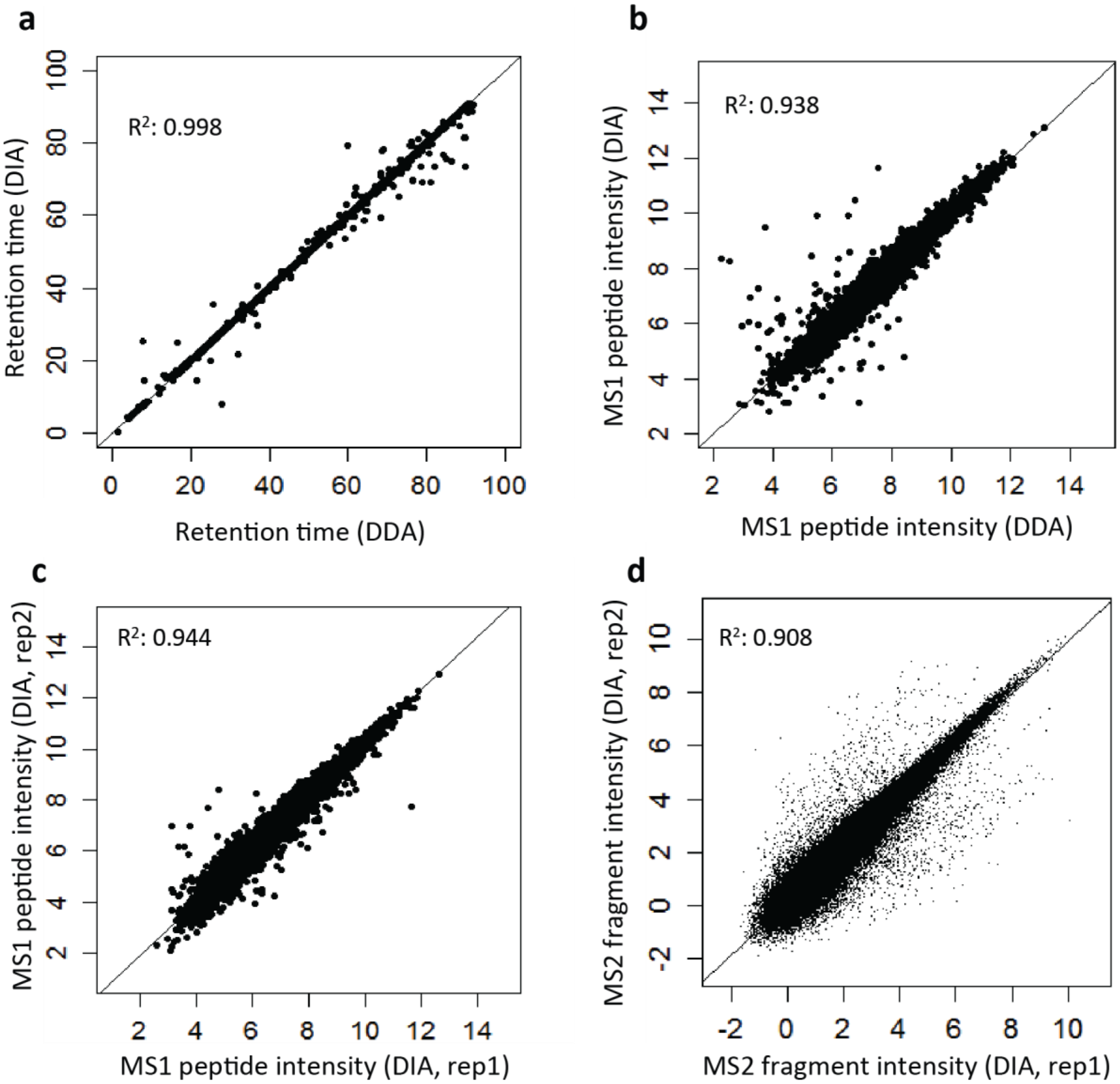


No. missed enz. cleavages (nmc)	
pos model	0.993 (nmc=0) 0.007 (1<=nmc<=2)
neg model	0.69 (nmc=0) 0.31 (1<=nmc<=2)

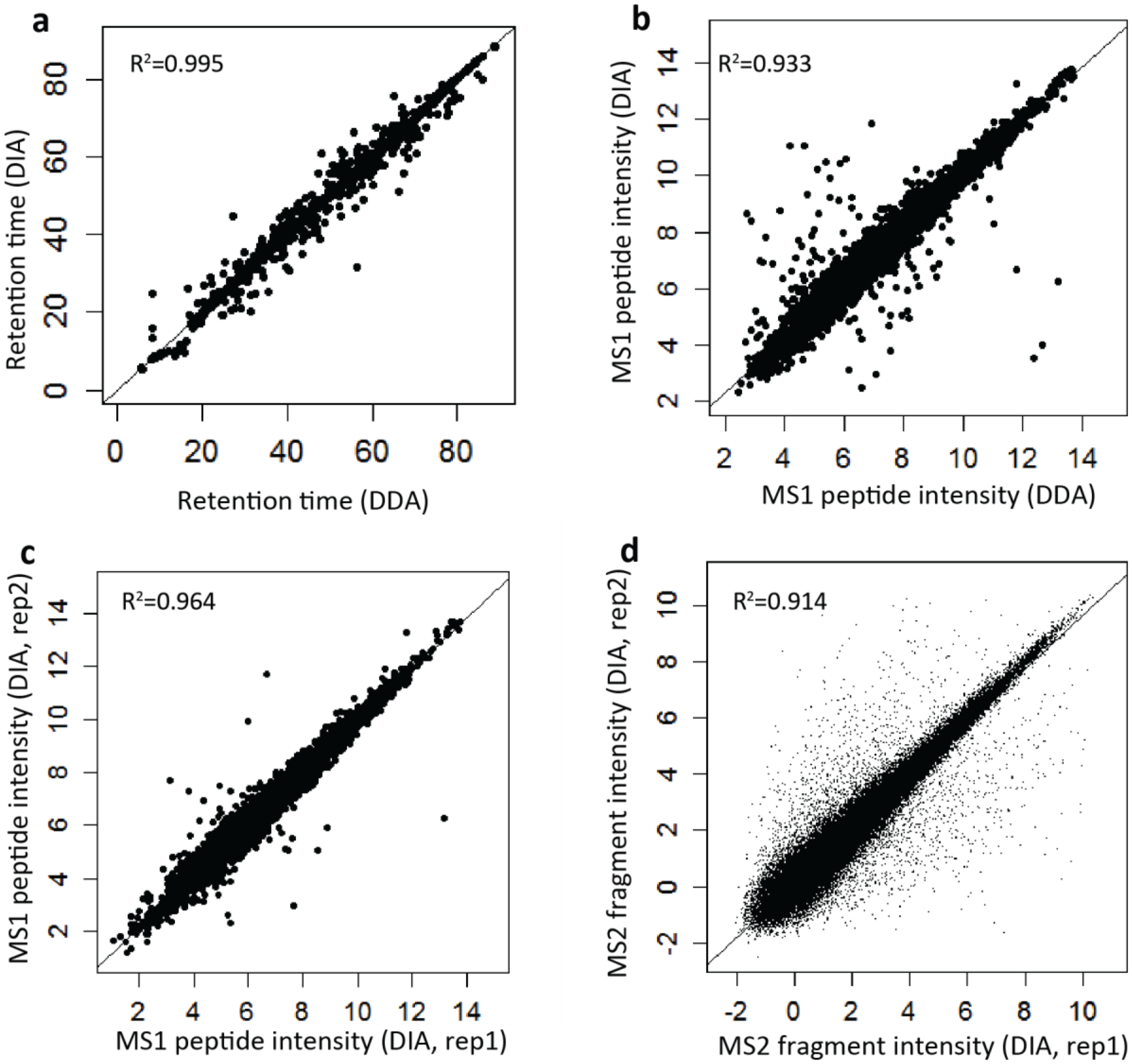
No. missed enz. cleavages (nmc)	
pos model	0.988 (nmc=0) 0.012 (1<=nmc<=2)
neg model	0.64 (nmc=0) 0.36 (1<=nmc<=2)

No. missed enz. cleavages (nmc)	
pos model	0.987 (nmc=0) 0.013 (1<=nmc<=2)
neg model	0.63 (nmc=0) 0.37 (1<=nmc<=2)

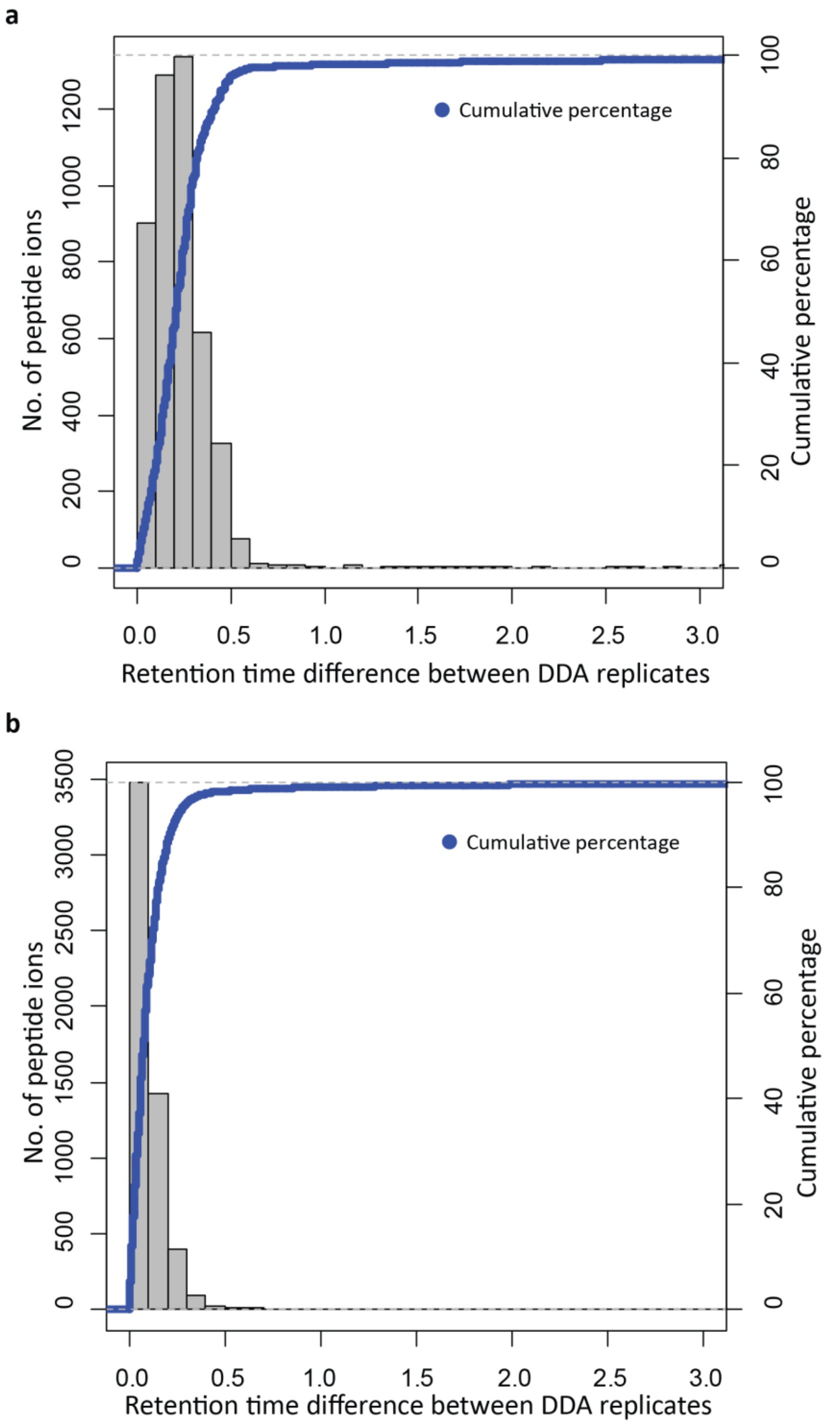
Supplementary Fig. 26. PeptideProphet analysis of X! Tandem search results using DDA and DIA pseudo MS/MS data. Shown are model distributions learned by PeptideProphet in the analysis of X! Tandem search results (doubly charged peptide ions) for one replicate of the human cell lysate data. Left panels: mass accuracy distributions. Right panels: the distributions of the discriminant database search scores (computed from the X! Tandem expect scores). Red and blue curves represent the models learned by PeptideProphet for correct and incorrect identifications, respectively. Also shown are the distributions for the number of missed cleavages parameter (nmc) among correct and incorrect identifications. **(a)** DDA data; **(b)** DIA data, QT=1 pseudo MS/MS spectra; **(c)** DIA data, QT=2 spectra. The learned distributions appear to be an accurate fit in both DIA and DDA data, demonstrating that the search results obtained using DIA pseudo MS/MS spectra can be satisfactory analyzed using PeptideProphet. The overall higher ratio of incorrect vs correct identification in the DIA QT=1 vs. DDA data (and similarly in DIA QT=2 vs. QT=1 data) simply reflects the higher number of pseudo MS/MS spectra extracted from the data compared to DDA data (and similarly, more noise in DIA QT=2 vs QT=1 data), which does not affect the accuracy of computed PeptideProphet probabilities or the subsequent FDR estimates.



Supplementary Fig. 27 Assessment of retention time and MS1 intensity reproducibility of identified peptide ions between DDA and DIA (SWATH) experiments. Human cell lysate data. (a) LC retention time (LC peak apex) for peptide ions identified in both DDA and DIA experiments. **(b)** MS1 intensities (monoisotopic peak intensities at LC peak apex) of peptide ions identified commonly by DDA and DIA. **(c)** Reproducibility of DIA MS1 peptide ion intensities between two replicates of DIA data. **(d)** Reproducibility of DIA MS2 fragment ion intensities (at the reconstructed LC peak apex) of peptide ions between two DIA replicates. Only matched *b*- and *y*-ion fragments were considered. Ion and fragment intensities are shown on $\log_2$ scale.



Supplementary Fig. 28 Assessment of retention time and MS1 intensity reproducibility of identified peptide ions between DDA and DIA (SWATH) experiments. *E. coli* cell lysate data. (a) LC retention time (LC peak apex) for peptide ions identified in both DDA and DIA experiments. **(b)** MS1 intensities (monoisotopic peak intensities at LC peak apex) of peptide ions identified commonly by DDA and DIA. **(c)** Reproducibility of DIA MS1 peptide ion intensities between two replicates of DIA data. **(d)** Reproducibility of DIA MS2 fragment ion intensities (at the reconstructed LC peak apex) of peptide ions between two DIA replicates. Only matched *b*- and *y*-ion fragments were considered. Ion and fragment intensities are shown on $\log_2$ scale.



Supplementary Fig. 29 Retention time differences for peptide ions commonly identified between DDA replicates. (a) *E. coli* cell lysate data **(b)** human cell lysate data.