Supporting Information

A three-stage search strategy combining database reduction and retention time filtering to improve the sensitivity of low-input and single-cell proteomic analysis

Wei Fang[1], Zhuokun Du[1], Linlin Kong[1], Guibin Wang[1], Yangjun Zhang[1], Weijie Qin[1,2*]

[1] State Key Laboratory of Proteomics, Beijing Institute of Lifeomics, National Center for Protein Sciences Beijing, Beijing Proteome Research Center, Beijing 102206, P.R. China

[2] College of Chemistry and Environmental Science, Hebei University, Baoding, 071002, China

* Correspondence: aunp_dna@126.com

Table of Contents

Parameters for the Conventional Search with Sequest HT and Sequest HT + Percolator

The test of Sequest HT engine was carried out on Proteome Discoverer Software (version 2.4, San Jose, CA). Raw files were searched against the human Uniprot database (release on 2021.06.29, 20386 entries). Cysteine carbamidomethylation was set as fixed modification (removed for samples without alkylation), whereas methionine oxidation and acetyl N-terminal were set as variable modifications. Trypsin/P was selected as the digestion enzyme with a maximum missed cleavage of two. The maximum peptide mass was set as 5000 and the mass tolerances of precursor ions and fragment ions were 5 ppm and 0.02 Da, respectively. The false discovery rate was set < 0.01 for both PSM and protein. The validation of PSM was performed using either Percolator (Sequest HT + Percolator search) or Target decoy (Sequest HT search). Other nodes such as "Spectrum Files", "Minora Feature Detector" and "Spectrum Selector" were used with default settings.

Parameters for the Three-stage Search with Sequest HT and Sequest HT + Percolator

Raw data was first searched using the same parameters as conventional search, except that the FDR were set < 0.4. Then, the identified proteins were used as a new database**.** The identified peptides with PEP < 0.025 and XCorr ≥ 1 were used to retrain the retention time (RT) prediction model AutoRT[1]. Secondly, the raw files were searched against the new database using the FDR setting < 0.1. Then the actual measured RTs of the peptides identified in the second search in different runs were calibrated using the AMRT[2] method and the predicted RTs were obtained by applying the retrained autoRT (modified peptides were excluded). Followingly, the ΔRT (difference between the measured and the predicted RT) for each peptide was calculated and used as a filtering criterion instead of PEP value to reduce the FDR. Briefly, the peptides were filtered according to the ΔRTs from large to small until the proportions of remaining peptides and proteins from decoy database were below 1% and the ΔRTs were limited to below 3 σ, in which σ stood for the standard deviation of the ΔRTs of confident peptides (score ≥ 20 and PEP < 0.025).

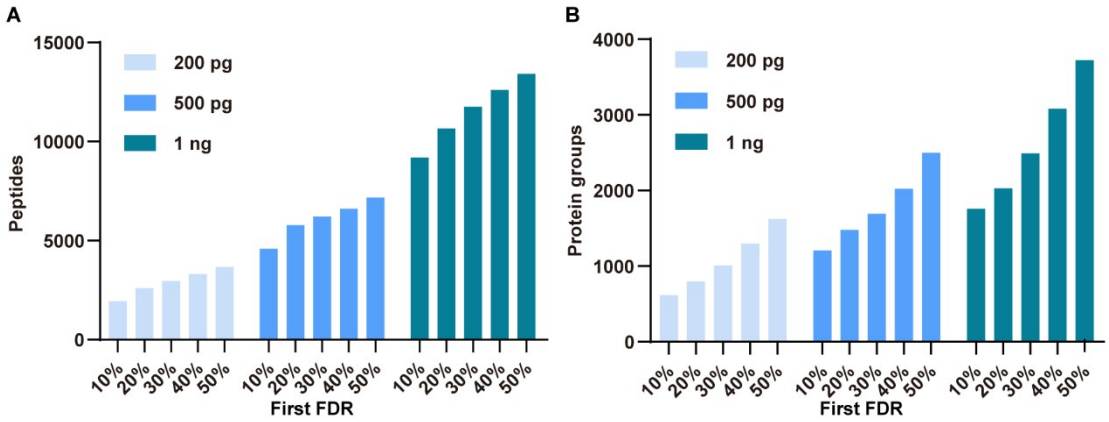

Fig. S1 Number of peptides (A) and protein groups (B) obtained using different FDR levels in the first search.

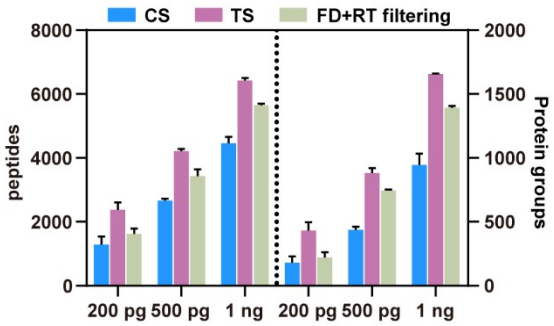

Fig. S2 Comparison of the number of peptides and proteins derived from 200 pg, 500 pg and 1 ng of peptides by conventional search, three-stage search and full database search with 10% FDR setting and ΔRT filtering. (CS: conventional search, TS: three-stage search, FD+RT filtering: full database search with 10% FDR setting and ΔRT filtering)

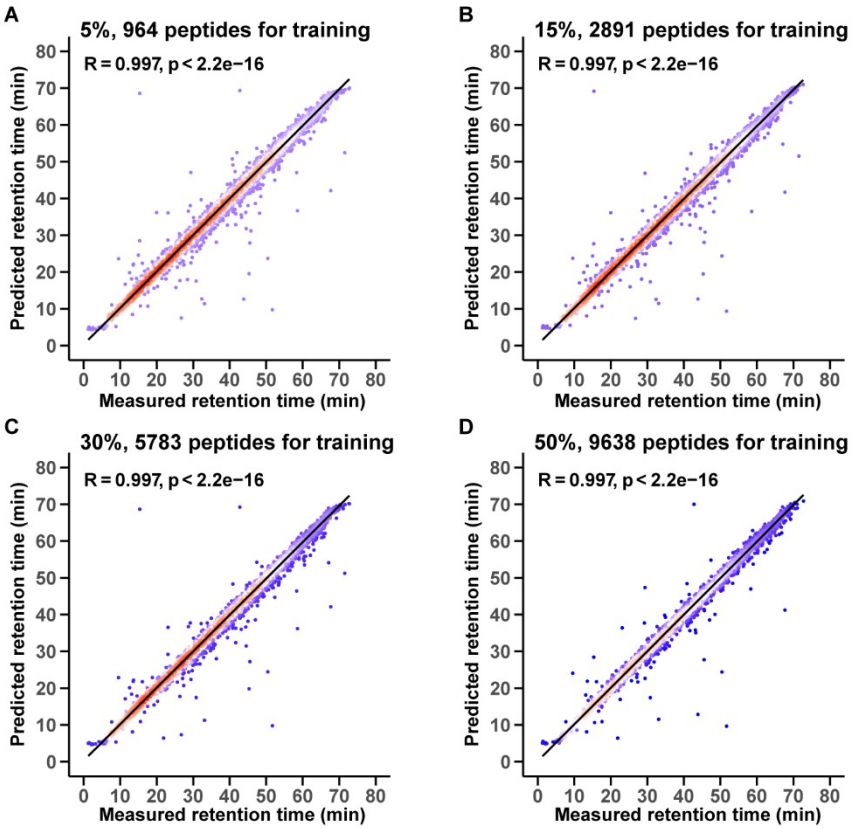

Fig. S3 Comparison of the Pearson coefficients of the predicted and measured retention times using different prediction models retrained by 5% (A), 15% (B), 30% (C) and 50% (D) of the total 19276 peptides. The remaining peptide datasets were used for testing, except for those used for training.

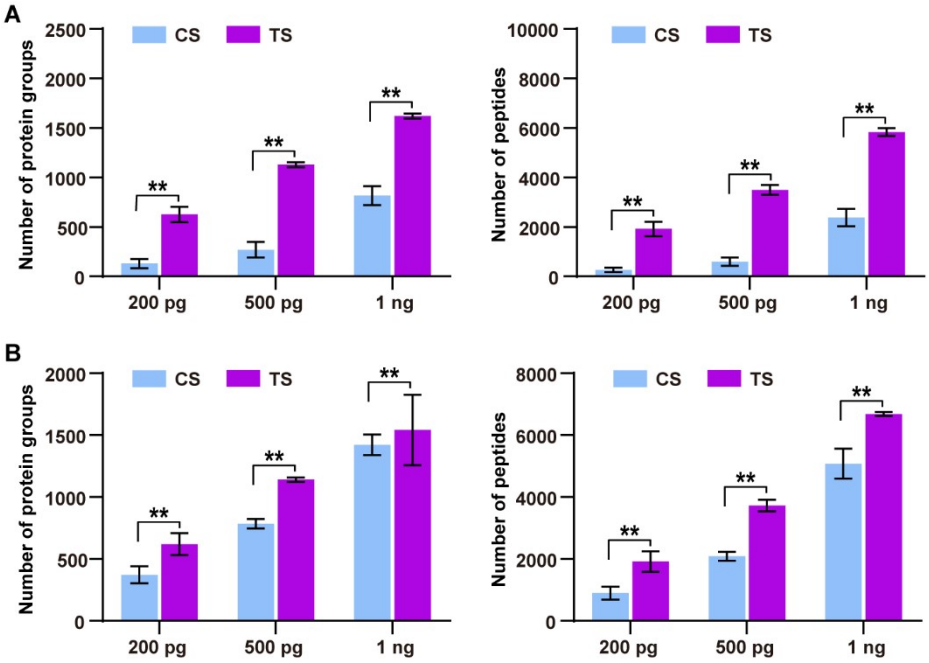

Fig. S4 Comparison of the number of protein groups and peptides identified by the conventional and the three-stage searches with the engines Sequest HT (A) and Sequest HT + Percolator (B). (CS: conventional search, TS: three-stage search, The ** symbol indicates a statistically significant increase ($P < 0.01$), N = 3)

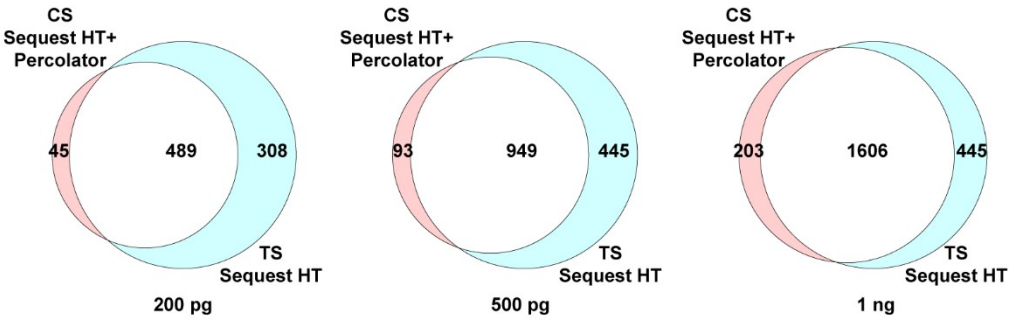

Fig. S5 Venn diagrams illustrating the overlap between Sequest HT + Percolator with the conventional search and Sequest HT with the three-stage search for protein groups derived from 200 pg – 1 ng of peptides. (CS: conventional search, TS: three-stage search)

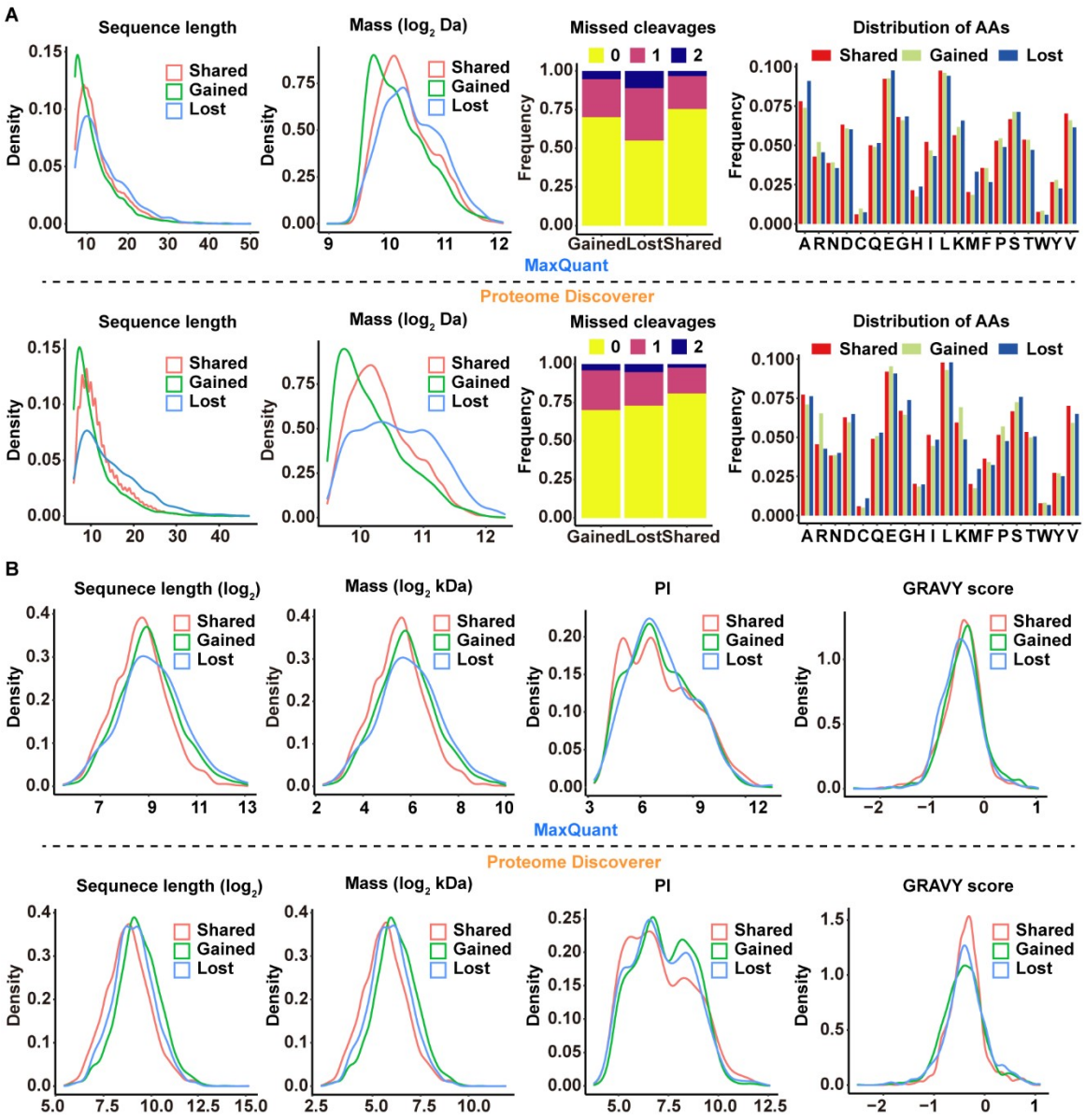

Fig. S6 Characteristic analysis of peptides and proteins gained and lost in TS compared to CS. (A) Kernel density estimation curve displaying the distribution of sequence length, mass, number of missed cleavages and AAs (amino acids) of peptides gained and lost in MaxQuant and Proteome Discoverer. (B) Kernel density estimation curve displaying the distribution of sequence length, mass, PI (isoelectric point) and GRAVY score of proteins gained and lost in MaxQuant and Proteome Discoverer. (CS: conventional search, TS: three-stage search)

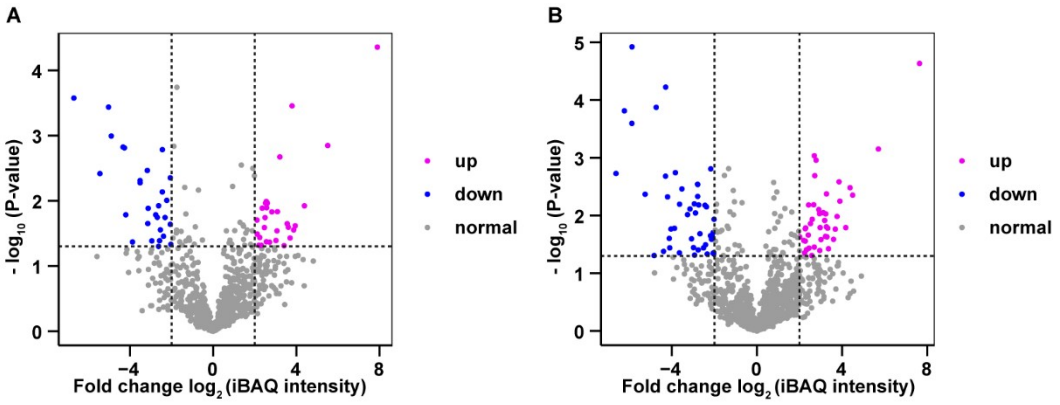

Fig. S7 Volcano plots displaying the differences in protein abundance between single interneurons and single motor neurons. (A) Volcano plot based on protein groups quantified by the conventional search. (B) Volcano plot based on protein groups quantified by the three-stage search. (Red dots represent proteins enriched in interneurons and blue dots represent proteins enriched in motor neurons, N = 3).

Table S1. ΔRT comparison of datasets with different sample sizes, gradient times and flow rates.

Sample	Gradient time (min)	Flow rate (nL/min)	ΔRT (min)	σ (min)	ΔRT/σ
500 ng (A)	75	300	1.539	0.513	3.00
50 ng (A)	75	300	1.602	0.534	3.00
10 ng (A)	75	300	1.128	0.376	2.74
1 ng (A)	75	300	0.669	0.236	2.83
500 pg (A)	75	300	0.693	0.286	2.42
200 pg (A)	75	300	0.861	0.287	3.00
100 cells (A)	75	300	1.110	0.379	2.93
10 cells (A)	75	300	0.806	0.359	2.25
1ng (D)	80	250	1.552	0.517	3.00
500 pg (D)	80	250	1.284	0.428	3.00
250 pg (D)	80	250	1.351	0.450	3.00
single cell (B)	160	20	3.600	1.200	3.00
100 cells (C)	170	12	3.612	1.338	2.70

A: datasets from our laboratory, B: datasets from PXD019515, C: datasets from PXD020669, D: datasets from PXD024017.

Table S2. Estimation of FDR for the three-stage search based on the results of the target-decoy strategy.

Sample	FDR_40%_10%				FDR_40%_10% _ΔRT Filtering				Peptide FDP [a] (%)	Protein FDP [a] (%)	Peptide FDP [b] (%)	Protein FDP [b] (%)
	Decoy peptides	Decoy proteins	Target peptides	Target proteins	Decoy peptides	Decoy proteins	Target peptides	Target proteins				
500 ng (A)	626±11	411±4	38,280±475	5,941±30	56±6	54±6	33,187±489	5,478±21	0.2	1.0	0.2	1.0
50 ng (A)	473±25	342±13	24,410±914	4,746±74	37±5	36±4	21,664±780	4,254±54	0.2	0.8	0.2	0.9
10 ng (A)	379±21	266±11	16,899±861	3,644±63	19±2	19±2	12,954±461	3,030±37	0.1	0.6	0.2	0.6
1 ng (A)	202±16	150±12	8,704±230	2,082±29	16±4	16±4	6,623±19	1,606±19	0.2	1.0	0.3	1.1
500 pg (A)	129±17	103±10	4,822±223	1,432±20	11±3	11±3	3,564±160	1,072±18	0.3	1.0	0.3	1.1
200 pg (A)	78±16	56±11	2,204±346	823±80	5±1	5±1	1,722±268	593±56	0.3	0.8	0.3	1.0
100 cells (A)	339±43	212±16	15,926±1469	2,794±148	24±2	23±3	14,157±1384	2,477±155	0.2	1.0	0.2	1.0
10 cells (A)	92±14	75±9	4,848±1278	1,302±195	10±2	10±2	3,623±983	1,002±173	0.3	1.0	0.3	1.1
Single-cell (B)	43±8	39±6	3,152±386	1,098±101	5±2	5±2	2,540±306	925±89	0.2	0.5	0.2	0.6
100 cells (C)	114±2	98±3	5,926±379	1,689±38	12±1	12±1	5,115±352	1,414±43	0.2	0.8	0.3	0.9
1 ng (D)	183±6	160±5	10,205±57	2,756±20	21±4	20±4	9,248±51	2,456±16	0.2	0.8	0.2	0.9
500 pg (D)	183±28	145±16	7,092±597	2,190±125	13±4	13±4	6,258±500	1,911±109	0.2	0.7	0.2	0.7
250 pg (D)	125±31	106±24	5,018±1221	1,687±304	12±5	11±5	4,433±1100	1,437±281	0.3	0.9	0.3	0.9
FDR	NA	NA	NA	NA	NA	NA	NA	NA	0.2	0.8	0.2	0.9

"Before ΔRT Filtering" referred to the results of the second search. “After ΔRT Filtering” meant the above results were further filtered with ΔRT. The FDP[a] was calculated using the formula: FDP = N_{decoy}/N_{target} × 100%. The FDP[b] was calculated using the formula: FDP = (N_{decoy}+1)/N_{target} × 100% as proposed in the literature[3] (He et al. https://arxiv.org/abs/1501.00537). (N = 3 – 6, A: datasets from our laboratory, B: datasets from PXD019515, C: datasets from PXD020669, D: datasets from PXD024017)

Table S3. Comparison of FDPs at peptide level between the conventional and the three-stage searches by mixing the yeast and human databases.

Sample	Conventional search (FDR: 1%)			Three-stage search (Before ΔRT Filtering)		Three-stage search (After ΔRT Filtering)		
	Yeast Peptides	Human peptides	Y/H (%)	Yeast Peptides	Human peptides	Yeast Peptides	Human peptides	Y/H (%)
500 ng (A)	20±3	29,644±511	0.1	242±14	35,912±369	35±4	32,642±395	0.1
50 ng (A)	16±3	17,497±1351	0.1	198±17	22,807±872	24±4	21,295±762	0.1
10 ng (A)	10±3	9,910±761	0.1	137±11	15,685±779	17±1	14,221±562	0.1
1 ng (A)	14±2	3,642±308	0.4	115±5	7,534±114	11±2	6,080±110	0.2
500 pg (A)	11±1	1,656±78	0.7	91±14	4,024±230	12±1	3,463±182	0.4
200 pg (A)	4±2	699±181	0.6	58±10	1,763±270	7±2	1,555±227	0.5
100 cells (A)	12±2	12,322±2,073	0.1	111±28	14,617±1,606	19±4	12,854±2,419	0.1
10 cells (A)	10±4	2,024±572	0.5	77±17	4,029±1120	18±3	3,148±894	0.6
Single-cell (B)	5±1	1,724±261	0.3	37±10	2,394±326	6±1	1,977±260	0.3
100 cells (C)	11±1	2,010±94	0.5	90±8	4,433±201	12±2	3,945±181	0.3
1 ng (D)	8±2	7,244±70	0.1	99±12	9,050±100	18±2	8,388±84	0.2
500 pg (D)	4±1	4,468±474	0.1	70±15	6,237±574	9±2	5,694±498	0.2
250 pg (D)	3±2	2,760±658	0.1	42±14	3,891±945	7±1	3,592±885	0.2
Average	NA	NA	0.3	NA	NA	NA	NA	0.3

In the three-stage search, the yeast database was mixed with the reduced human database to be searched against in the second step to estimate the FDP more rigorously. "Before ΔRT Filtering" referred to the results of the second search. “After ΔRT Filtering” meant the above results were further filtered with ΔRT. Y/H ratio is calculated as the ratio of the number of peptides identified from the yeast database to those identified from the human database. (N = 3 – 6, A: datasets from our laboratory, B: datasets from PXD019515, C: datasets from PXD020669, D: datasets from PXD024017)

Table S4. Comparison of FDPs at protein level between the conventional and the three-stage searches by mixing the yeast and human databases.

Sample	Conventional search (FDR: 1%)			Three-stage search (Before ΔRT Filtering)		Three-stage search (After ΔRT Filtering)		
	Yeast Proteins	Human Proteins	Y/H (%)	Yeast Proteins	Human Proteins	Yeast Proteins	Human Proteins	Y/H (%)
500 ng (A)	18±3	5,249±38	0.3	180±13	5,785±27	30±5	5,495±36	0.5
50 ng (A)	15±2	3,778±117	0.4	164±7	4,503±63	22±3	4,239±51	0.5
10 ng (A)	10±3	2,475±43	0.4	121±6	3,337±42	16±1	3,071±36	0.5
1 ng (A)	9±2	1,071±52	0.8	98±7	1,782±36	10±1	1,508±10	0.7
500 pg (A)	11±1	669±22	1.6	86±12	1,162±34	10±1	997±27	1.0
200 pg (A)	4±2	310±62	1.3	53±8	615±55	6±2	515±39	1.2
100 cells (A)	11±2	2,232±241	0.5	94±23	2,584±178	19±4	2,354±232	0.8
10 cells (A)	6±2	620±125	1.0	61±16	1,085±181	10±3	903±169	1.1
Single-cell (B)	5±2	663±79	0.8	36±9	908±110	6±1	793±94	0.8
100 cells (C)	11±0	799±24	1.4	86±6	1,402±31	9±3	1,220±25	0.7
1 ng (D)	8±2	2,030±20	0.4	84±10	2,505±14	18±2	2,300±18	0.8
500 pg (D)	4±1	1,409±83	0.3	63±13	1,926±131	9±2	1,774±109	0.5
250 pg (D)	3±2	991±180	0.3	41±13	1,340±254	7±1	1,231±240	0.6
Average	NA	NA	0.7	NA	NA	NA	NA	0.7

In the three-stage search, the yeast database was mixed with the reduced human database to be searched against in the second step to estimate the FDP more rigorously. "Before ΔRT Filtering" referred to the results of the second search “After ΔRT Filtering” meant the above results were further filtered with ΔRT. Y/H ratio is calculated as the ratio of the number of proteins identified from the yeast database to those identified from the human database. (N = 3 – 6, A: datasets from our laboratory, B: datasets from PXD019515, C: datasets from PXD020669, D: datasets from PXD024017)

Table S5. Comparison of the candidate peptides of the 52 low-quality spectra in the CS and the TS and the resulting candidate peptides after ΔRT filtering.

MSMS scan number	Candidate peptides in CS, ΔRT (min) was indicated in the brackets. The remaining candidate peptides after ΔRT filtering were labeled in yellow.	Candidate peptides in TS, ΔRT (min) was indicated in the brackets. The remaining candidate peptides after ΔRT filtering were labeled in yellow.
1214	LGEHNIK (2.27), KPSPEPR (0.90), AVEHINK (0.42), GILHQDK (2.40)	AVEHINK (0.42), KPSPEPR (0.90)
1520	VDEEQMK (0.06), VDCEIDK (1.88)	VDEEQMK (0.06)
1698	VAQDLCK (0.18), VAINCEK (1.27)	VAQDLCK (0.18)
1728	ALHQCNK (4.76), AICDHVR (0.20)	AICDHVR (0.20)
2128	ERDKEEGK (8.62), HNQELHGR (7.25), HLVYESDK (0.17), MFHLPMSK (28.20), GARCTVNGR (6.722), IEGTQADTR (3.87)	HLVYESDK (0.17)
2398	GTDYQLSK (0.27), ASSSLDGFK (7.38)	GTDYQLSK (0.27)
2572	ATAPQTQHVSPMR (0.75), TDKAEVVNGYEAK (2.04), ENAEVTYSLLER (31.27)	ATAPQTQHVSPMR (0.75)
2592	LPEPTTR (0.31), LGPEIER (5.64)	LPEPTTR (0.31)
2627	TDSDIIAK (0.27), SLDIDTAK (5.73)	TDSDIIAK (0.27)
2693	VAQPTITDNK (0.37), VYNVTYTVK (13.64), LNATYYITK (17.77)	VAQPTITDNK (0.37)
2796	QVEKVVDK (6.13), NIQLSLEK (18.75), SGEVLVNVK (9.98), VQTLIENK (2.96), VQTEVLQK (0.27)	SGEVLVNVK (9.98), VQTEVLQK (0.27)
2996	LSLGGYAK (10.34), TVGQLYK (0.43), ISANLYK (4.67), VTFNLSK (13.21)	TVGQLYK (0.43)
3126	EAGVVAQAR (5.77), AGEVFIHK (0.34)	AGEVFIHK (0.34)
3311	TITSSYYR (0.30), TIYFFGDK (29.98)	TITSSYYR (0.30)
3390	WSPVQSVEK (8.75), QPDSGISSIR (0.25)	QPDSGISSIR (0.25)
3434	ASLTGTPSR (7.31), ASLENSLR (0.11), IASTASSPR (10.79)	ASLENSLR (0.11)
3468	IISDKQR (12.27), LSKENIR (9.27), ISGTVNIR (0.62), LSERNLK (10.11)	ISGTVNIR (0.62)
3565	EFSGNPIK (0.74), EEWAKTK (10.30)	EFSGNPIK (0.74)
3566	IYDLFNR (24.15), LNQYFQK (0.60), IPWSFYK (30.53), LEGDHTIR (9.78), YTAAVPYR (2.50)	LEGDHTIR (9.78), LNQYFQK (0.60)
3607	SAPTSPCDQEIK (4.45), HQEGEIFDTEK (0.01)	HQEGEIFDTEK (0.01)
3673	HPDADSLYVEK (0.14), SQGSGNEAEPLGK (8.69), HDEAFSTEPLK (3.60)	HPDADSLYVEK (0.14)
3960	NVVLQYGFK (21.54), IAQITGPPDR (0.35)	IAQITGPPDR (0.35)
3989	IVVAMAK (0.10), LVAMAVK (2.66)	IVVAMAK (0.10)
4073	AQASPSEENK (17.79), EQNNDALEK (14.75),	DGYNYTLSK (0.75)

	DGYNYTLSK (0.75)	
4137	VVGLEGSDK (8.01), VMEHFIK (0.68)	VMEHFIK (0.68)
4245	LVGGTTPGK (13.21), IVEEALR (4.47), VLDVVER (0.65), LVEAIER (4.20)	LVGGTTPGK (13.21), VLDVVER (0.65)
4270	EIEFLPSR (18.57), TGIDLGTTGR (0.24), LGNSFVPEK (2.35)	TGIDLGTTGR (0.24)
4361	NLETPLCK (0.39), ARQANDTAK (21.9)	NLETPLCK (0.39)
4728	IIHTGEKPYK (15.32), IIYTGEKPHK (14.24), IINNTENLVR (0.58)	IIHTGEKPYK (15.32), IINNTENLVR (0.58)
4919	WMPPEAFLEGIFTSK (43.51), NSNLVGAAHEELQQSR (0.65)	NSNLVGAAHEELQQSR (0.65)
4948	VPASLKEK (17.60), VPTISINK (0.01), VDLVAQVK (1.48)	VPTISINK (0.01), VDLVAQVK (1.48)
4958	LEELTMDGAK (0.14), ELRCQCIK (16.28)	LEELTMDGAK (0.14)
4998	QPSQGPTFGIK (0.30), ERLEAASQQK (20.77), QHEVDKLYK (16.50)	QPSQGPTFGIK (0.30)
5408	SLFGQVLK (18.78), IGQGYLIK (0.17)	IGQGYLIK (0.17)
5420	LLEIDISSNK (12.55), LLEFFGHLR (26.44), LLLENDSLSK (4.69), ILRERDSSR (24.28), LLSNDEVTIK (0.29)	LLSNDEVTIK (0.29)
5552	TSGTTAAPRVK (22.57), LGSNAGNKSLK (22.38), DKAVLNSVSR (15.27), AYPAPLTSIR (8.73), AKNTGVSVGQK (23.20), AKEKSEIQR (24.88), DSKLTHLFK (2.44), VLQLYPNNK (0.23), AEWLNKTVK (7.78), SKNLTDAIAR (10.69)	VLQLYPNNK (0.23)
5868	QELQSLK (18.64), LEQNTIK (23.98), YPDPLIK (0.73), AVQASIEK (22.69)	YPDPLIK (0.73)
6034	LLGTKTCK (24.11), LIGEYGLR (0.33)	LIGEYGLR (0.33)
6148	YCNYASKGTAR (25.21), ADGYEPPVQESV (0.29), YRSSDSSFWR (8.86)	ADGYEPPVQESV (0.29)
6210	FYWRPHCR (12.01), FYSVNVDYSK (0.18)	FYSVNVDYSK (0.18)
6267	CAGTVEVEIQR (10.73), CAGNEDIITLR (0.21)	CAGNEDIITLR (0.21)
6446	QADKVWR (24.82), GAWSNVLR (0.37)	GAWSNVLR (0.37)
6693	DFQAIADVIGNK (20.96), HRSDLIEHQR (29.77), LGGEVSCLVAGTK (0.66)	LGGEVSCLVAGTK (0.66)
7076	GIHPTIISESFQK (0.64), QSLDVPLREGTNK (12.15)	GIHPTIISESFQK (0.64)
7816	FPQLCKFCDVR (2.39), FPQLDSTSFANSR (0.09)	FPQLDSTSFANSR (0.09)
12070	GVPQIEVTFDIDANGILNVSAVDK (0.59),	GVPQIEVTFDIDANGILNVSAVDK

	QGFLPPLNVNSHPPISDINVNNK (9.23)	(0.59)
12109	ILIIGGSIANFTNVAATFK (0.61), LIQSEVALNDLHLTKQK (24.70)	ILIIGGSIANFTNVAATFK (0.61)
2780	LGVIEDHSNR (0.20), VLYMDKENK (0.56)	LGVIEDHSNR (0.20)
2972	ELEEELK (2.85), EIEELEK (0.59), EEELLEK (1.36), ELEEIEK (0.29), EEEELLK (2.89)	ELEEIEK (0.29)
6120	IIVGSFMGYLR (31.48), MPPAEKASRIR (22.38), ERDRVLPSQR (23.60), LTLTAVDGGSPPK (0.18), GLHNVVYGIQR (4.11), SLPTLHERFR (6.14), SAIPIGGGSRGAGR (16.34), LQNVNRDIQR (21.89), AVLIPKDDQEK (13.19), FDRPALPANVR (2.53), ILEDNSIPQVK (0.09)	ILEDNSIPQVK (0.09), MPPAEKASRIR (22.38)
7126	LWNTLGVCK (0.56), WERPFEVK (8.58), FSPGLPGYPR (0.62)	LWNTLGVCK (0.56)
12532	DKATGEVLGQFYLDLYPREGK (7.47), DIAFLEAQIYEYVEILGEQR (0.09), VKQSPGTKLCHGDSELTSGLLAT (28.91), QDQIHSSIVSTLLALMDGLDSR (0.55), AKMASVPVYCLCRLPYDVTR (15.27)	DIAFLEAQIYEYVEILGEQR (0.09)

References

(1) Wen, B.; Li, K.; Zhang, Y.; Zhang, B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nat. Commun.* **2020**, *11*, 1759.

(2) Silva, J. C.; Denny, R.; Dorschel, C. A.; Gorenstein, M.; Kass, I. J.; Li, G.-Z.; McKenna, T.; Nold, M. J.; Richardson, K.; Young, P. Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* **2005**, *77*, 2187-2200.

(3) He, K.; Fu, Y.; Zeng, W.-F.; Luo, L.; Chi, H.; Liu, C.; Qing, L.-Y.; Sun, R.-X.; He, S.-M. A theoretical foundation of the target-decoy search strategy for false discovery rate control in proteomics. *arXiv preprint arXiv:1501.00537* **2015**.