Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics

Supplementary information

Ryan Peckner, Samuel A. Myers, Alvaro Sebastian Vaca Jacome, Jarrett D. Egertson, Jennifer G. Abelin, Michael J. MacCoss, Steven A. Carr, and Jacob D. Jaffe

1 Supplementary Notes

1.1 Constructing the reference spectra library matrix

Let S be an acquired MS/MS spectrum from the DIA run. S is analyzed using only a subset of the provided spectral library, since there are physical constraints on the possible presence of a given library member in S.The set of library members used to analyze S is determined by the following conditions (where L denotes a candidate library precursor):

1. The m/z ratio of L must lie inside the precursor isolation window from which S was acquired.
2. At least five of the m/z ratios of the peaks of the spectrum of L must appear as m/z ratios of peaks S.
3. If the library includes retention time information, and the library retention times are directly comparable to those in the DIA experiment (as is the case if e.g. the library was generated from DDA runs of the same samples on the same instrument, or both the library and acquired spectra have had their retention times normalized), then the library retention time for L must be no more than five minutes greater or less than the time of the scan (this time window can be omitted or adjusted by the user).

While retention time information in the library is optional, it both speeds the analysis by limiting the set of precursors considered for each scan and improves the quality of the results, and so is highly encouraged to be included in cases that the library and DIA spectra are gathered in comparable timeframes.

For each MS2 scan S, the m/z coordinates of the peaks of the library spectra are then binned with the m/z coordinates of the peaks of S in the following way. Let $L_1, L_2, \ldots, L_m$ be the library spectra that satisfy the three criteria laid out above. Each spectrum L_j may be considered as a set of (m/z, intensity) pairs $(x_{jk}, y_{jk}), k = 1, \ldots, \ell_j$. Let $x_1, x_2, \ldots, x_n$ be the m/z coordinates of the peaks of S. Then we set

$$L_{ij} = \begin{cases} y_{jk} & \text{if there exists } k \leq \ell_j \text{ such that} \\ & \qquad |x_{jk} - x_i| < \delta x_{jk} \qquad \text{for } 1 \leq i \leq n \text{ and } 1 \leq j \leq m. \\ 0 & \text{if not} \end{cases}$$

L_{ij} is well-defined as long as δ is sufficiently small, since the triangle inequality implies that we can't have both $|x_{jk} - x_i| < \delta x_{jk}$ and $|x_{jk} - x_{i'}| < \delta x_{jk}$ for $i \neq i'$ as long as $|x_i - x_{i'}| > \delta x_{jk}$ for every pair of m/z coordinates of distinct peaks of S.

Since the above only accounts for peaks of library spectra that actually appear in the acquired spectrum S, an extra term $L_{(n+1)j}$ is constructed, whose value is the sum of the intensities of all peaks of L_j which do not appear in S (i.e. are not sufficiently close to any of the peaks of S, as measured by the parameter δ). This serves to penalize library spectra that have spurious peaks not appearing in the observed data. The library spectrum L_j is then represented as the intensity vector $L_j = (L_{ij})_{1 \leq i \leq n+1}$. In order to interpret the coefficients produced by Specter as total ion intensities (§**??**), each library spectrum vector is then normalized so that the sum of its entries equals one. This yields m normalized intensity vectors $L_1, \ldots, L_m \in \mathbb{R}^{n+1}$, and the library reference spectra matrix is the $(n+1) \times m$ matrix whose columns are these vectors:

$$\begin{aligned} L &= \left[L_1^T \, L_2^T \cdots L_m^T \right] \\ &= (L_{ij})_{\substack{1 \leq i \leq n+1 \\ 1 \leq j \leq m}}. \end{aligned}$$

1.2 Finding the optimal combination

Let S be an MS2 scan from the DIA experiment, represented as a vector of n intensity values, and let $L \in \mathbb{R}^{n \times m}$ be the corresponding matrix of normalized reference spectra constructed above. In order to account for peaks of library spectra not matching peaks in S, as described above, we append a zero to the end of this vector so that it has length $n+1$. This extra zero serves to penalize the linear contributions of library spectra that have peaks with significant intensities that aren't present in S. Let $L \in \mathbb{R}^{(n+1) \times m}$ be the corresponding matrix of normalized reference spectra constructed above. Our aim, as stated in Equation (1), is to find the nonnegative linear combination of the columns of L (the normalized library spectra) that best explains S, i.e. is closest to it in terms of Euclidean distance. Some peaks of S may not be close to any of the peaks of the reference spectra, as determined using the mass accuracy δ, and these may be discarded from the analysis as they don't affect

the determination of the optimal linear combination (see below for proof of this), i.e. we project the spectrum S to the linear span of the library spectra prior to analysis.

With these unnecessary peaks removed, the optimal linear combination of the reference spectra is determined as the solution of the corresponding nonnegative least squares problem:

$$\text{Find } c \in \mathbb{R}^m \text{ which minimizes } \|S - Lc\|^2 \text{ subject to } c \geq 0,$$

where the condition $c \geq 0$ means $c_j \geq 0$ for $j = 1, \ldots, m$, and the Euclidean norm is used. The objective function $f(c) = \|S - Lc\|^2$ for this convex optimization problem need not be strictly convex in general, which raises the possibility of the existence of multiple solutions to this nonnegative least squares problem. However, in most real cases, the library spectrum matrix is overdetermined (there are more fragments in the acquired DIA spectrum than there are library spectra) and has linearly independent columns (no library spectrum can be written as a linear combination of other library spectra), resulting in a strictly convex optimization problem that does have a unique global minimum. We find this global minimum by means of gradient descent applied to the objective function f, implemented via the Python package `cvxopt` v. 1.1.9.

To see that the projection of S to the linear span of the library spectra has no effect on the coefficients calculated for each library spectrum, observe we may write the residual as

$$\begin{aligned}\|S - Lc\|^2 &= \sum_{i=1}^{n} |S_i - (Lc)_i|^2 \\ &= \sum_{i=1}^{n} \left| S_i - \sum_{j=1}^{m} c_j L_{ij} \right|^2 .\end{aligned}$$

If for some i it happens that x_i is not close to any of the peaks of the reference spectra, then we have $L_{ij} = 0$ for $j = 1, \ldots, m$. Let

$$B = \{1 \leq i \leq n : |x_i - x_{jk}| \geq \delta x_{jk} \text{ for } 1 \leq j \leq m, 1 \leq k \leq \ell_j\}.$$

Then we see that

$$\|S - Lc\|^2 = \sum_{i \in B} |S_i|^2 + \sum_{i \notin B} \left| S_i - \sum_{j=1}^{m} c_j L_{ij} \right|^2 .$$

Since $\sum_{i \in B} |S_i|^2$ is a fixed contribution to the objective function f, we have

$$\nabla f = \nabla \left(f - \sum_{i \in B} |S_i|^2 \right)$$

and therefore projecting S to those of its coordinates whose indices are not in B (i.e. removing all peaks from S that aren't sufficiently close to any peak of any library spectrum) has no effect on the determination of the c_j.

1.3 Chromatographic peak scores

The previous section describes how Specter computes a coefficient, interpreted as a total ion intensity, for each spectral library member within each acquired MS2 spectrum of the DIA run. We can view the series of coefficients associated to a given library member as a time series $c_1, c_2, \ldots, c_r$, where r is the number of acquired MS2 spectra (note that most of these coefficients will be zero). The precursor is initially considered to be identified by Specter if there exist indices $1 \leq s < p < e \leq r$ such that

1. $c_i > 1$ for $s \leq i \leq e$
2. $c_{i+1} > c_i$ for $s \leq i \leq p - 1$
3. $c_{i+1} < c_i$ for $p \leq i \leq e - 1$
4. $p \geq s + 2$ and $e \geq p + 2$.

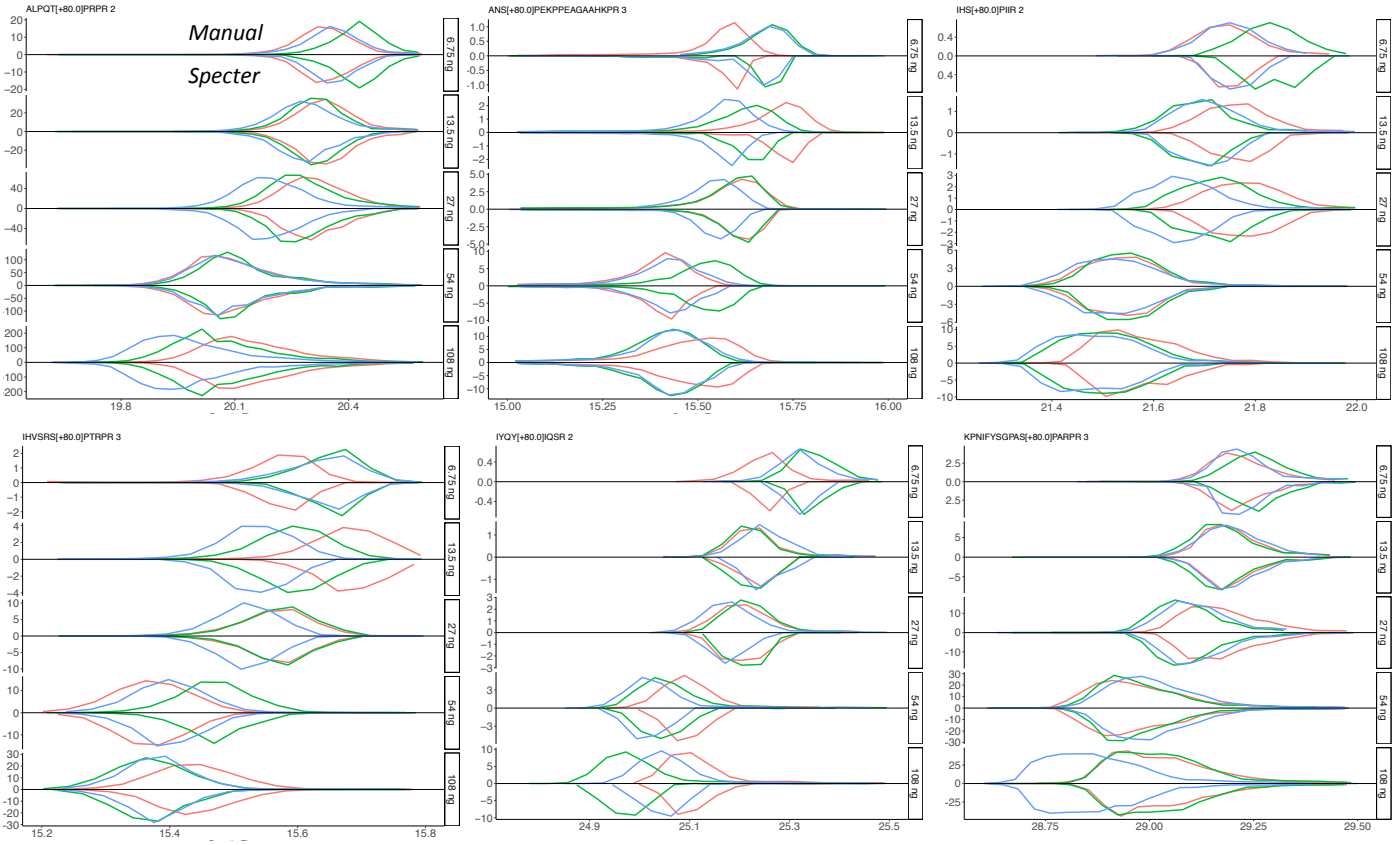
In other words, for the precursor to be initially considered identified, its sequence of Specter coefficients must contain a peak of at least five consecutive values greater than 1. Quantification and scoring of an initially identified precursor is based on the highest peak in the sequence of Specter coefficients. Let $c_s, c_{s+1}, \ldots, c_e$ be the sequence of Specter coefficients comprising the

highest peak for a given precursor. Let $\bar{c}$ be their mean and σ their standard deviation. Then the four scores associated to the precursor are

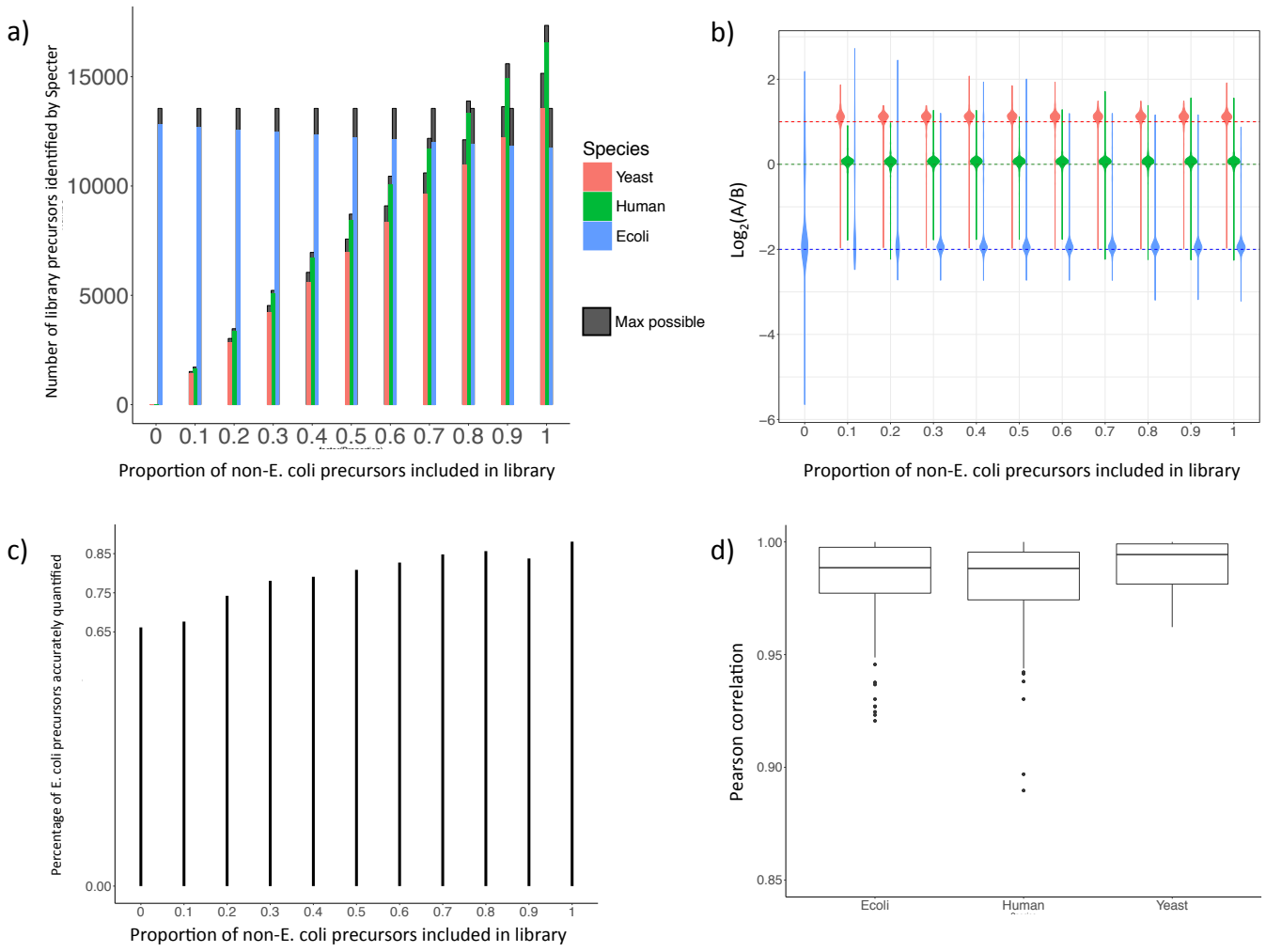
$$s_1 = \max_{s \le i \le e} c_i$$
$$s_2 = \sigma^2$$
$$s_3 = \frac{1}{2\sigma^3} \sum_{i=s}^{e} (c_i - \bar{c})^3$$
$$s_4 = \frac{1}{3\sigma^4} \sum_{i=s}^{e} (c_i - \bar{c})^4 - 3.$$

These are the maxima and standardized central moments (variance, skewness and kurtosis) of the coefficients comprising the highest peak. 3 is subtracted from s_4 in order to achieve a kurtosis of zero for the Gaussian. These scores measure the asymmetry and peakedness of the highest chromatographic peak relative to an ideal Gaussian, and are used for linear discriminant analysis to separate target and decoy library members and establish FDRs (§ 2.2.1).

2 Supplementary Figures and Tables

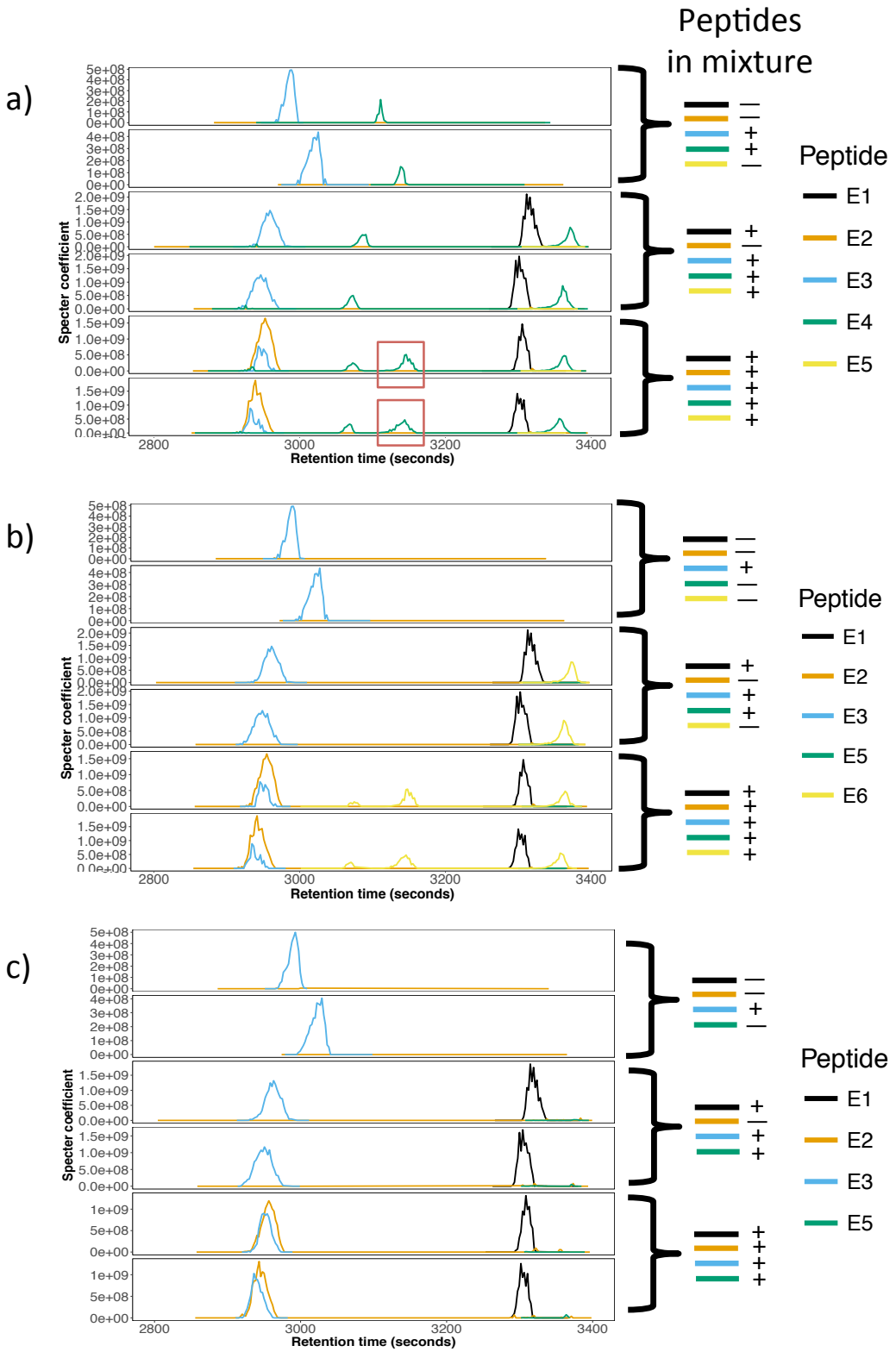


Supplementary Figure 1. Comparisons of manually determined and Specter chromatograms for synthetic phosphopeptides in HEK293T background, cf. Figure 2. $n = 3$ technical replicates for each comparison.

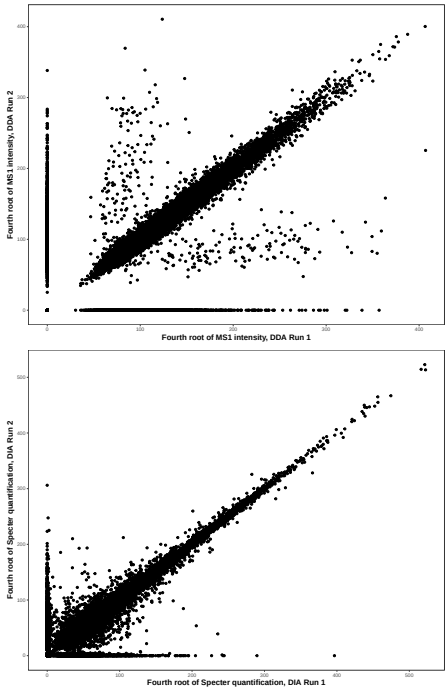


Supplementary Figure 2. Analysis of the LFQ Bench dataset (Section 2.6) using incomplete spectral libraries. The original spectral library containing all yeast, human and E. coli precursors was sequentially subsetted to contain all of the E. coli precursors, plus $0\%, 10\%, 20\%, \ldots, 90\%, 100\%$ of the yeast and human precursors. a) Numbers of library precursor identifications for each species and proportion considered. b) Violin plots showing distributions of log-ratios of Specter quantifications between the two LFQ Bench samples for each species and each proportion of non-E. coli precursors included. Dashed lines indicate the expected log-ratio values for each species. Violin plots represent distributions of $n = 12409, 15327, 18257, 21150, 23948, 26826, 2968, 32374, 35083, 37797, 40594$ quantification ratios for proportion $0, 0.1, ..., 0.9, 1$, and are calculated by means of kernel density estimation. c) Percentage of E. coli precursors accurately quantified (meaning the log-ratio is within 0.5 of the expected value) for each proportion of non-E. coli precursors included. In this context, inaccurate quantification likely corresponds to misidentification of E. coli precursors as human or yeast. d) Distributions of correlations between Specter quantifications for all pairs of distinct proportions of non-E. coli precursors (for example, one point in the distribution would be the correlation between the quantifications for all precursors of a given species identified when using the subsetted libraries containing all E. coli precursors and either 10% or 70% of the non-E. coli precursors). Each boxplot represents a distribution of $n = 100$ correlations; whiskers extend from the minimum to maximum of the correlation values except where outliers are indicated, and the elements of the boxes indicate the 25th, 50th, and 75th percentiles of each distribution.

E1: LPLVLANGQIR E4: LPVLAVNGQIR
E2: LPVVLANGQIR E5: LPVLANVGQIR
E3: LPVLVANGQIR E6: LPVLANGVQIR



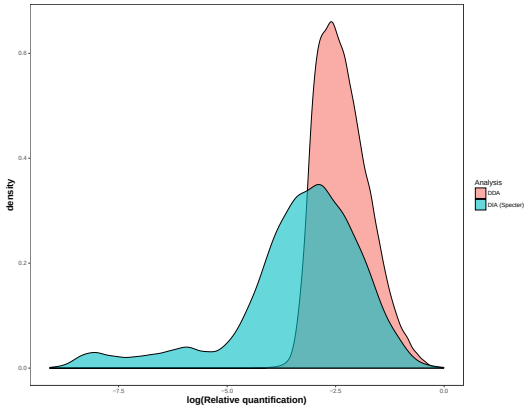
Supplementary Figure 3. Analysis of the similar synthetic peptides dataset using incomplete spectral libraries. a) When E6 is removed from the spectral library, E4 is identified in its place. b) When E4 is removed from the spectral library, no other peptide is identified in its place. c) When both E4 and E6 are removed, no other peptides are identified in place of either of these two.



(a)

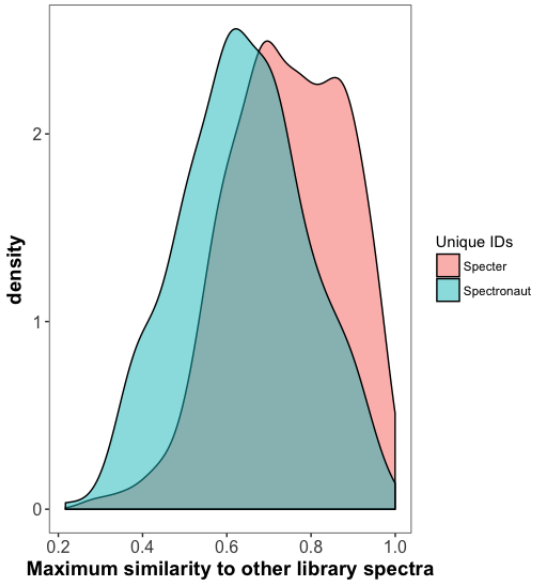
Replicate pair	R^2 (DDA)	R^2 (DIA)
(1, 2)	0.73	0.98
(1, 3)	0.73	0.99
(2, 3)	0.71	0.98

(b)



(c)

Supplementary Figure 4: DIA quantification by Specter offers superior reproducibility and a wider dynamic range in comparison to DDA.
(a) *Top:* Fourth root of MS1 intensities of library precursors across replicate DDA runs of an E. coli digest. *Bottom:* Fourth root of Specter quantitations of library precursors across replicate DIA runs of the same sample.
(b) Table of values of R^2 between replicates for precursor quantifications by DDA vs DIA.
(c) Distributions of logarithms (base 10) of relative quantifications (ratio of absolute quant to maximum) of all precursors.



Supplementary Figure 5. Distributions of maximum similarities (normalized dot products) between spectra of peptides uniquely identified by Specter or Spectronaut and all other library spectra from the LFQ Bench study.

Window	Center (m/z)
1	400.43
2	422.44
3	444.45
4	466.46
5	488.47
6	510.48
7	532.49
8	554.50
9	576.51
10	598.52
11	620.53
12	642.54
13	664.55
14	686.56
15	708.57
16	730.58
17	752.59
18	774.60
19	796.61
20	818.62
21	840.63
22	862.64
23	884.65
24	906.66
25	928.67
26	950.68
27	972.69
28	994.70
29	411.44
30	433.45
31	455.46
32	477.47
33	499.48
34	521.49
35	543.50
36	565.51
37	587.52
38	609.53
39	631.54
40	653.55
41	675.56
42	697.57
43	719.58
44	741.59
45	763.60
46	785.61
47	807.62
48	829.63
49	851.64
50	873.65
51	895.66
52	917.67
53	939.68
54	961.69
55	983.70
56	1005.71

Supplementary Table 1: Window centers for DIA experiments of §2.3.

Peptide	m/z	Isolation windows
GFSASSAR	391.6932	[389.43,411.43]
GFSANSAR	405.1987	[389.43,411.43],[400.43,422.43]
IVQDYLEK	504.2740	[499.48,521.48],[488.48,510.48]
TVQDYLEK	498.2558	[488.48,510.48],[477.47,499.47]
AVQDYLEK	483.2506	[477.47,499.47],[466.47,488.47]
LPLVLANGQIR	597.3719	[576.52,598.52],[587.52,609.52]
LPVVLANGQIR	590.3640	" "
LPVLVANGQIR	590.3640	" "
LPVLAVNGQIR	590.3640	" "
LPVLANVGQIR	590.3640	" "
LPVLANGVQIR	590.3640	" "

Supplementary Table 2: Precursor m/z ratios and isolation windows of synthetic peptides for the experiment of §2.3.

Peptide	Mixture 1	Mixture 2	Mixture 3
GFSASSAR	·	·	·
GFSANSAR		·	·
IVQDYLEK			·
TVQDYLEK		·	·
AVQDYLEK	·	·	·
LPLVLANGQIR		·	·
LPVVLANGQIR			·
LPVLVANGQIR	·	·	·
LPVLAVNGQIR	·	·	·
LPVLANVGQIR		·	·
LPVLANGVQIR			·

Supplementary Table 3: Design for the experiment of §2.3. A dot indicates that the peptide was spiked into the indicated mixture.

Species	Sample B intensity tertile	OpenSWATH	Spectronaut	PeakView	Skyline	DIAUmpire	Specter
Ecoli	1	1.02058	0.5679113	0.3590821	0.7023023	0.5270977	0.1601132
Ecoli	2	0.6218639	0.2988857	0.228552	0.3299477	0.4183653	0.160301
Ecoli	3	0.1925423	0.1618297	0.1144882	0.1262852	0.313052	0.1766875
Yeast	1	0.3260988	0.267417	0.1511932	0.2058005	0.2232352	0.2945058
Yeast	2	0.2087182	0.1869254	0.1479262	0.1643527	0.1807951	0.1971551
Yeast	3	0.1631462	0.1221844	0.1835488	0.1934923	0.2023442	0.1864905
Human	1	0.1346807	0.148201	0.1412818	0.1545666	0.2184005	0.1294194
Human	2	0.1257585	0.1124631	0.1278716	0.1304916	0.1739002	0.1347911
Human	3	0.1193357	0.1043191	0.1229887	0.1206303	0.1688108	0.1276729

Supplementary Table 4. Accuracy of relative peptide quantification (median absolute deviation from expected log2(A/B) values) for Specter and five other DIA analysis tools for high precision quantifications (CV across replicates < 10%) from LFQ Bench study data, grouped by tertiles of intensity in sample B.

Peptide	Amount (fmol)
ALGS[+80]PTKQLLPC[+57]EMAC[+57]NEK	3.39
VSMPDVELNLKS[+80]PK	55.77
LGPGRPLPTFPTSEC[+57]TS[+80]DVEPDTR	45
LPLVPES[+80]PRR	3.66
S[+80]FAGNLNTYKR	5.91
S[+122]DKPDMAEIEKFDK	7.5
S[+122]DKPDM[+16]AEIEKFDK	20.775
QDDS[+80]PPRPIIGPALPPGFIK	27
VYT[+80]HEVVTLWYR	46.71
KPNIFYSGPAS[+80]PARPR	37.47
AFGSGIDIKPGT[+80]PPIAGR	12
LAAPSVSHVS[+80]PR	11.4
QGSGRES[+80]PSLASR	22.185
IHVSRS[+80]PTRPR	12.24
TNPPTQKPPS[+80]PPMSGR	15.36
S[+80]PPAPGLQPMR	0.9
S[+80]PPAPGLQPM[+16]R	1.59
RPHS[+80]PEKAFSSNPVVR	4.59
SEVQQPVHPKPLS[+80]PDSR	2.58
NEEPVRS[+80]PERR	2.415
LFIIRGS[+80]PQQIDHAK	4.5
DRS[+80]SPPPGYIPDELHQVAR	135
LGM[+16]LS[+80]PEGTC[+57]K	2.565
LGMLS[+80]PEGTC[+57]K	8.325
RNS[+80]SEASSGDFLDLK	23.865
AAPEAS[+80]SPPASPLQHLLPGK	22.5
TPKDS[+80]PGIPPSANAHQLFR	13.56
RLS[+80]ESQLSFRR	14.715
SMS[+80]VDLSHIPLKDPLLFK	300
S[+80]PTGPSNSFLANMGGTVAHK	3.585
S[+80]LTNSHLEKK	2.67
SPDKPGGS[+80]PSASRR	2.535
IGPLGLS[+80]PK	750
A[+42]TTATMATSGS[+80]AR	12.045
A[+42]TTATM[+16]ATSGS[+80]AR	3.585
SSDQPLTVPVS[+80]PK	2.43
VLS[+80]PTAAKPSPFEGK	11.01
ETPHS[+80]PGVEDAPIAK	5.355
FYETKEESYS[+80]PSKDR	37.155
HLPS[+80]PPTLDSIITEYLR	30.615
RLS[+80]QSDEDVIR	8.115
ATS[+80]PVKSTTSITDAK	1.62
LENS[+80]PLGEALR	3.78
ISNLS[+80]PEEEQGLWK	18.105
RLIS[+80]PYKK	4.8
TLGRRDS[+80]SDDWEIPDGQITVGQR	22.5
IYQY[+80]IQSR	5.52
ANS[+80]PEKPPEAGAAHKPR	13.86
RLS[+80]LPGLLSQVSPR	60
IHS[+80]PIIR	3.375
DLVQPDKPAS[+80]PK	1.785
TFS[+80]LTEVR	4.215
S[+80]IQDLTVTGTEPGQVSSR	10.095
LHS[+80]APNLSDLHVVRPK	22.5
S[+122]DNGELEDKPPAPPVR	7.5
HAS[+80]PILPITEFSDIPR	15
LIPGPLS[+80]PVAR	3.21
LEVTEIVKPS[+80]PK	23.52
ST[+80]FHAGQLR	3.78
SNS[+80]LPHSAVSNAGSK	5.655

SPALKS[+80]PLQSVVVR	3.525
TPS[+80]IQPSLLPHAAPFAK	4.77
SQS[+80]PHYFR	2.28
HRPS[+80]PPATPPPK	2.91
S[+80]IPLSIK	32.13
VGS[+80]LDNVGHLPAGGAVK	2.73
SPS[+80]PAHLPDDPKVAEK	3.48
ANAS[+80]PQKPLDLK	4.62
VLS[+80]PLIIK	8.895
LLEDS[+80]EESSEETVSR	15.165
TQLWASEPGT[+80]PPLPTSLPSQNPILK	262.5
NDS[+80]WGSFDLR	35.37
KAYS[+80]FC[+57]GTVEYMAPEVVNR	225
KAYS[+80]FC[+57]GTVEYM[+16]APEVVNR	375
SLVGS[+80]WLK	44.25
SDS[+80]PENKYSDSTGHSK	2.445
SFS[+80]ADNFIGIQR	126.15
RRLS[+80]SLR	34.17
ANS[+80]FVGTAQYVSPELLTEK	300
QIT[+80]MEELVR	13.11
QIT[+80]M[+16]EELVR	18.63
YGS[+80]PPQRDPNWNGER	37.005
SFS[+80]SQRPVDR	6.24
SST[+80]PLPTISSSAENTR	18.6
AGS[+80]PDVLR	8.715
S[+80]LTAHSLLPLAEK	1.71
TEFLDLDNSPLSPPS[+80]PR	10.5
VDDDS[+80]LGEFPVTNSR	15.375
YLLGDAPVS[+80]PSSQK	7.275
ALPQT[+80]PRPR	83.445
TAPTLS[+80]PEHWK	17.37
LAS[+80]PELER	1.785
LQS[+80]EPESIR	2.175
LQTPNT[+80]FPKR	9.105
SLS[+80]LGDKEISR	1.635
S[+80]IEVENDFLPVEK	22.5

Supplementary Table 5: Amounts of synthetic phosphopeptides used for the lowest spike-in level of 6.75ng in the experiment of §2.2. The mixture is at an overall concentration of 4.5 ng/ul.

Compound Number	Compound	Concentration (μM)
1	**DMSO**	0.1
2	Selumetinib	5.5
3	PD0325901	4
4	Everolimus	0.1
5	vemurafenib	10
6	TG101348	8
7	Tofacitinib	0.4
8	**Pravastatin**	0.75
9	PD-0332991	0.5
10	Dinaciclib	4.5
11	**RO4929097**	3.5
12	BMS-906024	0.5
13	Verteporfin	5
14	vorinostat	1
15	SCH 900776	5.5
16	VX-970	0.1
17	losmapimod	0.25
18	PRI-724	1.7
19	dactolisib	1
20	afuresertib	2.5
21	BYL719	0.25
22	Pazopanib	40
23	**Nilotinib**	3
24	lenalidomide	1
25	AR A014418	3
26	IPI145	0.5
27	**staurosporine**	1
28	PS-1145	10

Supplementary Table 6: Drug treatments and concentrations for phosphoenriched PC3 samples. Bold font indicates that a DDA run of samples treated with the indicated perturbation was included in construction of the spectral library used for DIA.