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

Ryan Peckner[1], Samuel A Myers[1], Alvaro Sebastian Vaca Jacome[1], Jarrett D Egertson[2], Jennifer G Abelin[1,3], Michael J MacCoss[2], Steven A Carr[1] & Jacob D Jaffe[1]



**Mass spectrometry with data-independent acquisition (DIA) is a promising method to improve the comprehensiveness and reproducibility of targeted and discovery proteomics, in theory by systematically measuring all peptide precursors in a biological sample. However, the analytical challenges involved in discriminating between peptides with similar sequences in convoluted spectra have limited its applicability in important cases, such as the detection of single-nucleotide polymorphisms (SNPs) and alternative site localizations in phosphoproteomics data. We report Specter (https://github.com/rpeckner-broad/Specter), an open-source software tool that uses linear algebra to deconvolute DIA mixture spectra directly through comparison to a spectral library, thus circumventing the problems associated with typical fragment-correlation-based approaches. We validate the sensitivity of Specter and its performance relative to that of other methods, and show that Specter is able to successfully analyze cases involving highly similar peptides that are typically challenging for DIA analysis methods.**

Mass spectrometry with data-dependent acquisition (DDA) is the method of choice for large-scale discovery proteomics, but this technique is fundamentally limited in terms of reproducibility and comprehensiveness because of the stochastic nature of its data-gathering process[1], which inhibits the consistent detection of proteins across samples. Targeted strategies such as parallel reaction monitoring and selected reaction monitoring allow reproducible measurement of low-abundance analytes or observation of prespecified targets across multiple samples[2], but this gain in specificity comes at the cost of a vastly limited range of observable precursors. DIA is a newer approach that combines the reproducibility of selected reaction monitoring with the breadth of DDA by simultaneously fragmenting all precursors whose mass-to-charge ratios ($m/z$) fall into one of a small number of wide windows that traverse the entire $m/z$ range. This results in convoluted MS2 spectra whose fragment ion intensities may comprise contributions from multiple peptide precursors and that are far more complex to analyze than their DDA counterparts.

The challenges posed by DIA demand specialized software tools for downstream analysis[3–7]. Most available tools apply targeted methods that require a user-provided spectral library to define the search space of peptides (and, in turn, proteins) that can be identified and quantified in the acquired data. These tools for the most part score library members relative to acquired MS2 spectra on the basis of characteristics such as normalized dot product, fragment ion correlation, and chromatographic peak shape. Although these scores typically penalize assignments to library spectra whose annotated fragment b- or y-ions are judged to exhibit interferences, these methods do not rigorously account for the confounding effects of precursor cofragmentation, which limit the ability to distinguish precursors with shared spectral features. Alternatively, untargeted methods[4,5,8] deconvolute the data directly without the use of spectral libraries on the basis of the grouping of fragment ions with correlated elution profiles. This analysis implicitly discards fragments with significant interferences owing to their poor correlation with a precursor's elution profile. Although this approach is promising for the discovery of previously unobserved analytes, the fact that it uses no prior information as provided by a spectral library may lead to a high false negative rate with complex samples[9] and makes it more susceptible to missing data than targeted methods in attempts to quantify analytes across multiple conditions.

Here we describe Specter, an algorithm for the identification and quantification of spectral library members in DIA data. It recognizes and formalizes the fundamental distinction between DIA and DDA, namely, the cofragmentation of potentially large numbers of precursors, some of which may share fragment ion $m/z$ values. Specter is based on a mathematical formulation of the cofragmentation problem, which is then solved by means of linear algebra. In contrast to the usual approach involving the

[1]Proteomics Platform, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Department of Genome Sciences, University of Washington, Seattle, Seattle, Washington, USA. [3]Present address: Neon Therapeutics, Cambridge, Massachusetts, USA. Correspondence should be addressed to J.D.J. (jjaffe@broadinstitute.org).

RECEIVED 19 JUNE 2017; ACCEPTED 23 FEBRUARY 2018; PUBLISHED ONLINE 2 APRIL 2018; DOI:10.1038/NMETH.4643

detection of correlated chromatographic profiles of selected precursor fragment ions, spectral deconvolution takes place purely at the MS2 level and involves the entire sequence of $m/z$ coordinates and relative intensities of library spectra peaks. This allows for the direct calculation of extracted ion chromatograms of library precursors, which can then be visualized and analyzed via traditional chromatographic approaches[10].

This approach is neither spectrum-centric (it does not match acquired MS2 spectra to those in a database) nor peptide-centric (it does not assess the evidence for individual library members in the acquired data), in contrast to other existing methods. Rather, it is 'combination-centric', in that it identifies and quantifies the single combination of library spectra that best explains an acquired MS2 spectrum. Specter eliminates the need to reduce spectra to curated fragment ions, carries an intrinsically low false discovery rate, and is able to distinguish precursors with highly similar library spectra, such as those originating from SNPs or positional isomers in phosphoproteomics data. The linear algebraic framework establishes a meaningful notion of the quantification of a precursor in a single DIA MS2 spectrum independently from chromatographic information. Specter is able to analyze DIA-type data from any instrument and acquisition scheme (e.g., SWATH or MS$^E$), and accepts experimental data files in centroided mzML format and spectral libraries in Bibliospec's blib format[11]. Retention-time information in the library is optional, and retention-time normalization is not required, though it might improve the speed and quality of the results. Specter is built on the open-source distributed computing framework Apache Spark and is available as an open-source software tool at https://github.com/rpeckner-broad/Specter.

## RESULTS

### Specter is based on algebraic deconvolution of mixed spectra

Specter is based on the hypothesis that every MS2 spectrum $S$ acquired in the course of a DIA run is a linear combination of the spectra of the precursors cofragmented to acquire it, including the effects of biochemical noise, instrument error, and experimental variability in a peptide's fragmentation pattern (**Fig. 1**). This is because the total number of ions with a particular $m/z$ in $S$ is, in ideal terms, simply the sum of the number of ions with that $m/z$ contributed by each of the constituent precursors of $S$. Furthermore, the number of ions with a certain $m/z$ that are produced by any one of these cofragmented precursors is entirely determined by the precursor's fragmentation pattern (its pure spectrum) and abundance at the time when $S$ is acquired. This allows us to associate a sequence of 'Specter coefficients' to each spectral library precursor that quantifies that precursor's contributions to the acquired DIA MS2 spectra and serves as a calculated total ion chromatogram (**Supplementary Notes 1**–**4**). These algebraic coefficients are then analyzed further to determine the final identifications and quantifications of library members (Online Methods).

The recognition that mixed mass spectra are linear combinations of pure spectra is an established principle in gas chromatography–mass spectrometry[12,13] and, more recently, metabolomics[14,15]. However, so-called matrix methods for spectral deconvolution have for the most part not taken shape in usable software for gas chromatography–mass spectrometry applications[13], and to the best of our knowledge Specter is the first software tool to apply linear deconvolution to mass spectrometry proteomics data.

### Specter is as accurate as targeted manual analysis

To compare the quantitative performance of Specter to that of a traditional manual analysis, we spiked a mixture containing 85 synthetic phosphopeptides into a complex HEK293T cell lysate at five concentrations ranging from 6.75 ng to 108 ng (of total peptide mixture) per injection and then measured each spike-in sample in triplicate on a Q-Exactive Orbitrap HF using DIA (Online Methods). We applied Specter to each resulting data file, using a spectral library consisting of the 85 synthetic peptides and 29,248 HEK293T precursors acquired from DDA runs of each sample. Independently, we carried out a targeted manual analysis of the unprocessed data for the synthetic peptides in Skyline.

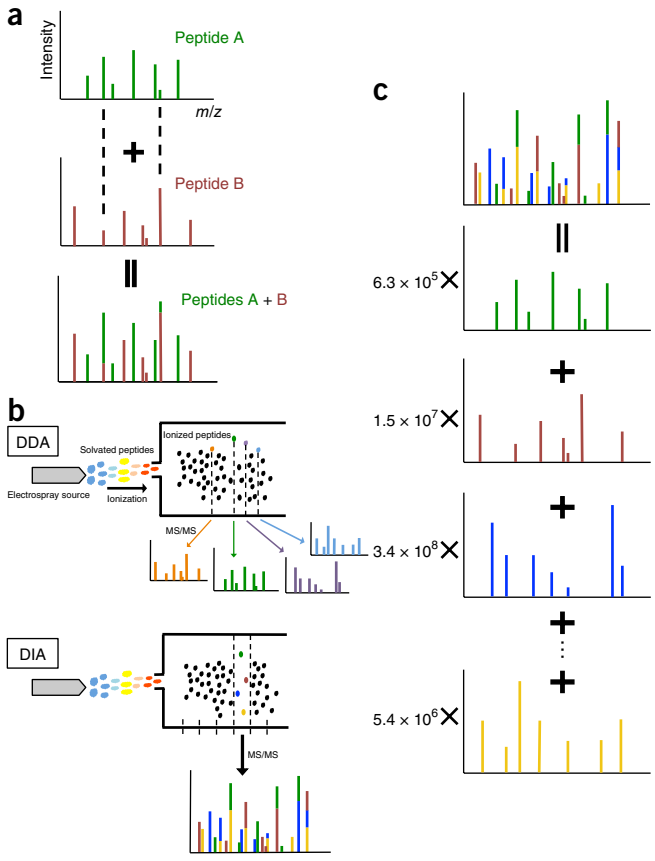


**Figure 1** | Specter uses linear algebra to formally deconvolute MS2 spectra derived from cofragmented precursors. (**a**) Mixed spectra are linear combinations of pure spectra. Pure spectra of two hypothetical precursors are shown. Their cofragmentation results in the combination of their pure spectra to yield a mixed spectrum containing fragment ion intensities from both precursors. Fragments with identical $m/z$ in the two pure spectra (indicated by dashed lines) lead to peaks in the mixed spectrum whose intensities are the result of contributions from both precursors. (**b**) In DDA, precursors are selected in decreasing order of abundance and fragmented separately to form MS2 spectra that typically represent a single precursor. In DIA, groups of precursors whose $m/z$ fall into the same wide window are fragmented simultaneously to form mixed MS2 spectra. (**c**) Specter finds the quantitative combination of library spectra that most closely matches the acquired DIA spectrum by linearly deconvolving the mixed spectrum into pure components from the library. The coefficient of each library spectrum is the total ion intensity of the corresponding precursor in the acquired spectrum.

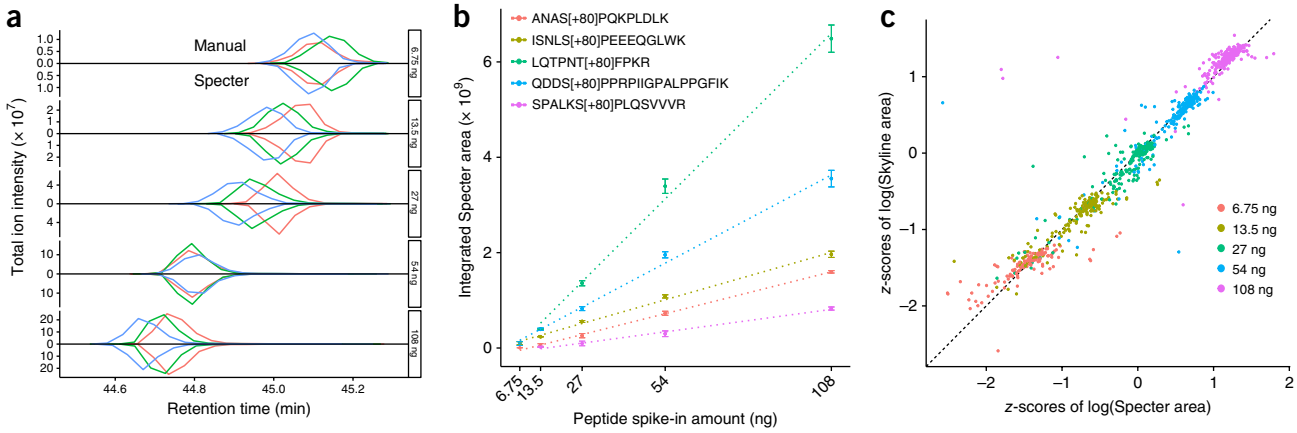


**Figure 2** | Total ion chromatograms calculated by Specter are as accurate as those from manual targeted analysis of DIA data in terms of both identification and quantification. Synthetic phosphopeptides were spiked into HEK293T lysate over a range of increasing concentrations, and each spike-in sample was analyzed in triplicate in DIA mode. (**a**) Total ion chromatograms determined by manual quantification in Skyline for the synthetic phosphopeptide VLS[+80]PLIIK, as well as those calculated by Specter, for the five peptide spike-in amounts. Each colored line for each concentration describes the total ion chromatogram for that precursor in a single replicate run, and chromatograms from each type of analysis are normalized to have the same maximum within each replicate. Note the increasing *y*-axis scales. (**b**) Mean ± s.e. of quantifications by Specter (area under the Specter chromatogram) across replicates for five of the synthetic peptides over the range of spike-in amounts, together with linear fits (dotted lines). (**c**) Comparison of *z*-scores of quantifications by Specter with those for manual quantification for all 85 synthetic peptides in all replicates and at all spike-in levels. The *z*-scores were calculated in such a way that the mean of the *z*-scores of each peptide across all spike-in levels and replicates is 0 and their s.d. is 1.



The elution profiles calculated by Specter agreed extremely well with the results of manual annotation in terms of both total ion intensities and the retention times at which they were identified (**Fig. 2a** and **Supplementary Fig. 1**). The quantifications derived by calculation of the area under the Specter chromatogram for each precursor (Online Methods) showed the expected linear increase with spike-in concentration (**Fig. 2b**) and agreed with manual quantifications, with a Pearson correlation of 0.96 between the *z*-scores of each method for all 85 synthetic peptides over all spike-in levels and replicates (**Fig. 2c**).

### The false discovery rate of Specter is intrinsically low

We adopted the target–decoy approach[16,17] to assess the false discovery rate of Specter. To test its robustness with acquired data, we augmented the focused spectral library mentioned above with 'decoys' from an *Escherichia coli* spectral library generated on the same instrument (no *E. coli* proteins were present in the sample). Of the 29,333 HEK293T and synthetic library precursors, 8,867 were cumulatively identified by Specter in the three replicate DIA runs of the 6.75-ng spike-in sample, whereas 297 of the 48,131 *E. coli* precursors were cumulatively identified, yielding an 'intrinsic' false discovery rate of 2% (**Fig. 3a** and Online Methods). This can be further reduced by means of linear discriminant analysis[18] (**Fig. 3b**, Online Methods, and **Supplementary Note 5**), at the cost of decreasing the number of on-target identifications.

### Specter is robust to spectral library incompleteness

Because Specter examines all of the precursors in a spectral library simultaneously, it might be susceptible to error when a spectral library lacks entries for a substantial number of the precursors likely to be present in a sample, as is commonly the case. We tested the robustness of Specter in such situations and found that its output was largely unaffected when precursors were removed individually or *en masse* from a spectral library (**Supplementary Figs. 2** and **3**).

### Specter distinguishes precursors with highly similar spectra

Spectral libraries often contain spectra that share a high number of peaks. Nonsynonymous SNPs in coding regions or alternative localizations for post-translational modifications (PTMs) can result in peptides whose spectra contain a paucity of discriminating fragment ions. Ambiguous shared features are typically deemed 'interferences' and excluded from consideration in analysis of DIA data. Such a strategy runs the risk of limiting biological insight, as differential expression of certain SNPs or PTMs may lie at the heart of disease phenotypes[19]. We assessed Specter's ability to distinguish

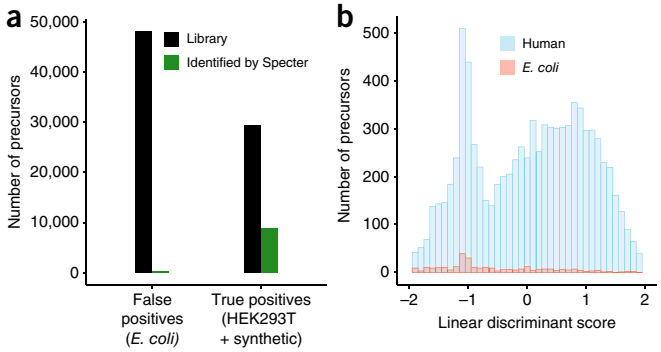


**Figure 3** | The false discovery rate of Specter is inherently <5%. (**a**) Specter cumulatively identified 30% of true positive HEK293T and synthetic library members and 0.6% of false positive *E. coli* library members in three replicates of the 6.75-ng spike-in sample. (**b**) Distributions of linear discriminant scores based on Specter coefficients of identified precursors from both libraries mentioned in **a**.

between extremely similar library spectra with large numbers of shared fragments by comparing Specter results to results obtained by use of the normalized dot product, the tool used by most other targeted DIA analysis methods to quantify spectral similarity.

We carried out DIA analysis of mixtures of synthetic peptides whose sequences differed by only a single amino acid or by the transposition of a pair of adjacent amino acids. We analyzed three families of synthetic peptides, each consisting of precursors whose spectra were highly similar and whose *m*/*z* (in charge state +2) fell into the same isolation windows for DIA (**Fig. 4** and **Supplementary Tables 1** and **2**). We then prepared a series of three mixtures, spiking a random set of members of each family into both an *E. coli* lysate digest and a neat background for each mixture (**Supplementary Table 3**). We analyzed each spike-in via DIA mass spectrometry in duplicate. For Specter analysis we used a spectral library consisting of 48,131 *E. coli* precursors together with the spectra of the synthetic peptides.

Specter distinguished between the synthetic precursors in each family despite their extremely similar library spectra, the cofragmentation of several precursors in each group (indicated by overlapping chromatograms), and the presence of the complex *E. coli* background (**Fig. 4**; results for the neat background were similar but are not shown). Note that library retention-time information was not used in any way for this analysis. In contrast, the normalized dot product analysis was unable to disambiguate the members of a group of six peptides with extremely similar spectra (**Fig. 4c**). Specter correctly identified all but one peptide (LPVLANVGQIR) in all runs; we had the prior expectation that this unidentified peptide would be problematic (**Supplementary Note 6**).

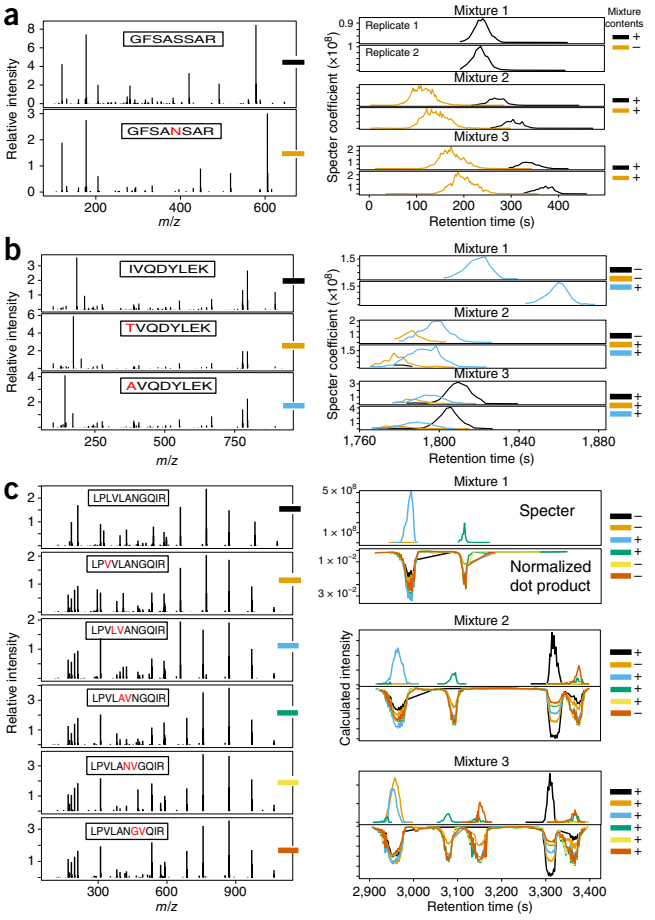


**Figure 4** | Specter chromatograms of groups of synthetic peptides with highly similar spectra. Members of each of three groups of highly similar peptides (spectra shown on the left in each panel) were chosen at random to be spiked into both an *E. coli* lysate and neat background in each of 12 DIA runs. Chromatograms for each peptide calculated by Specter in each replicate run of each mixture are shown on the right in each panel. Actual combinations of peptides in runs are indicated in the color keys, with each color corresponding to a distinct member of each group (only data from runs in the *E. coli* background are shown). (**a**) A single amino acid substitution. (**b**) Two unique substitutions at the N-terminal position, creating identical y-ion series for all family members. (**c**) A larger family consisting of substitutions and transpositions at various positions in the sequence. Comparison of chromatograms for each peptide in this family as calculated by Specter versus the normalized dot product for each mixture (data for only one replicate per mixture are shown).

### Distinguishing positional isomers in phosphoproteomics data

The analysis of positional isomers (peptides with identical amino acid sequences but with PTMs in different positions) is challenging for DDA approaches depending on where fragmentation spectra are sampled during elution, and it has only recently been explored for DIA data[20,21]. We analyzed 84 DIA runs of a set of phosphopeptide-enriched samples obtained from PC3 prostate cancer cells subjected to a panel of 28 kinase-pathway inhibitors in biological triplicate on a Thermo Q-Exactive Plus HF. For Specter analysis we used a 12,546-member phosphopeptide library constructed from ten DDA runs of the phosphoproteome of PC3 cells subjected to a subset of the kinase-inhibitor perturbations. This library contained 176 sets of unambiguous positional isomers, as determined by Spectrum Mill's variable modification location score.

Of the 12,546 phosphopeptides, an average of 2,218 were identifiable per run, with 327 identified in all 84 runs considered (**Fig. 5a**). On average, each phosphosite was identified in 41 of the 84 runs considered. This compares favorably with findings from a recent study of a large sample cohort conducted via SWATH-MS (which for our purposes is equivalent to DIA), in which the average reproducibility of phosphosite identifications was 2.6% (ref. 21) (it should be noted that only 24 phosphosites were considered in that study and that the conditions being compared were more heterogeneous than is the case here, as the study samples were derived from different patient samples). Among the 176 sets of positional isomers in the spectral library, at least one member of each set was identified in 31 of the 84 runs on average, and both members were identified in 17 of the runs on average (**Fig. 5b**).

Specter identified both of the positional isomers GYYS[+80]PYSVSGSGSTAGSR (S4613, in reference to the position of the phosphorylated serine on the underlying protein) and GYYSPYSVS[+80]GSGSTAGSR (S4618) in 75 of the 84 runs, and for most of these cases the isomers' elution profiles overlapped in retention time (**Fig. 5c**). We used Specter to disambiguate the similar spectra (**Fig. 5d**) and to quantify the ratio of the ion currents of the two isomers (**Fig. 5e**).

The peptide GYYSPYSVSGSGSTAGSR is a constituent of the cytoskeletal cross-linking protein plectin, located in the last of six highly homologous repeat domains forming plectin's C terminus. The sequence SPYS in this peptide is a known binding motif for CDK1, and it has been shown by site-directed mutagenesis that CDK1 phosphorylates plectin somewhere in repeat domain 6

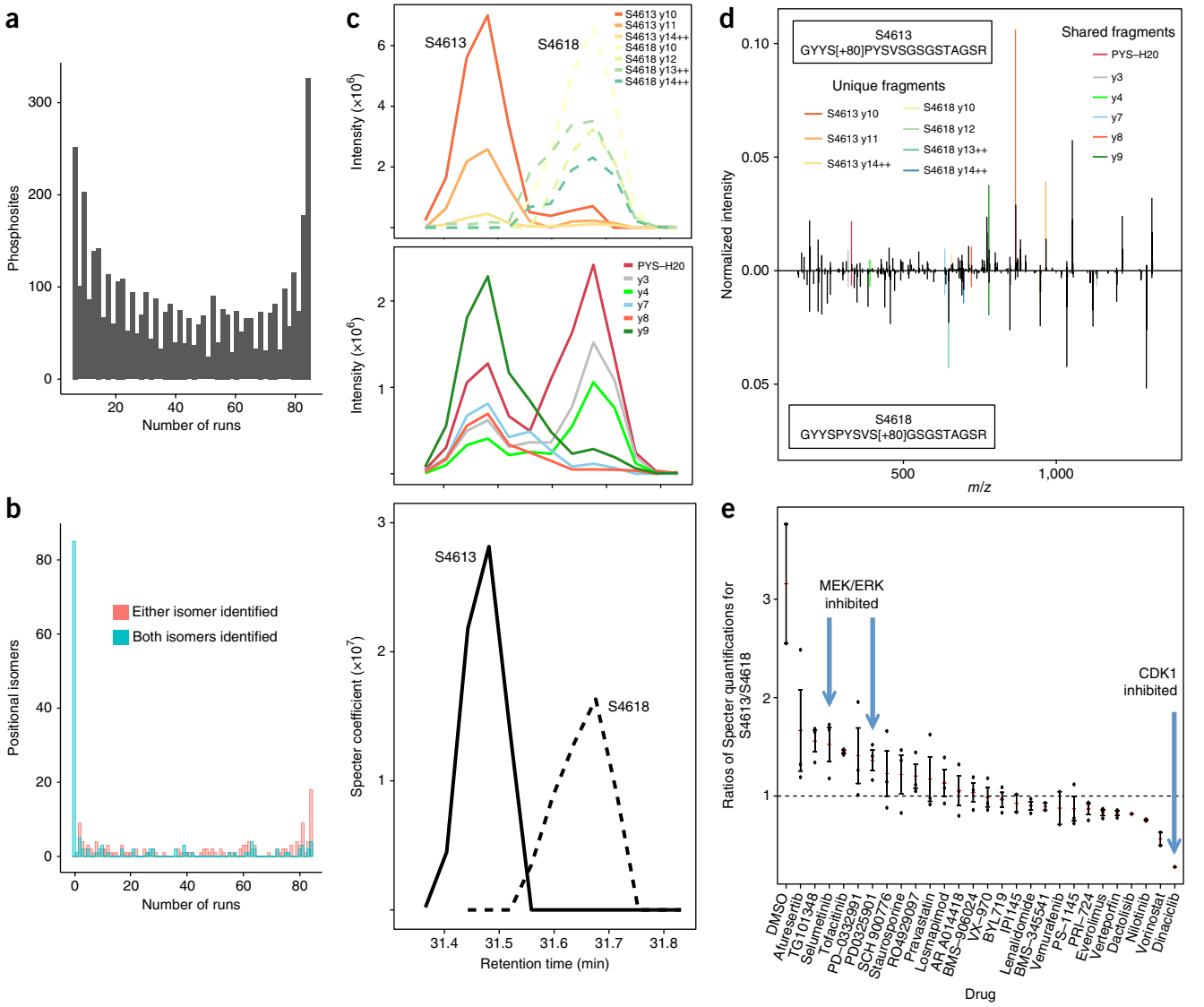


**Figure 5** | Specter distinguished close positional isomers with overlapping chromatographic profiles in a phosphoproteomics data set. (**a**) Distribution of numbers of phosphosite identifications across the 84 runs (excluding phosphosites from the library that were not identified in any of the runs). (**b**) Distributions of numbers of individual and pairwise identifications of positional isomers across the 84 runs (excluding isomers from the library that were not identified in any of the runs). (**c**) Top: raw extracted ion chromatograms for the unique fragments from each of the positional isomers in a DIA run of PC3 cells treated with TG101348. Middle: extracted ion chromatograms for the shared fragments from the isomers. Bottom: Specter chromatograms for each of the positional isomers. (**d**) The spectra of the positional isomers. Panels **c** and **d** represent data derived from single samples. (**e**) Ratios of the quantifications of the isomers by Specter (S4613/S4618) across 28 chemical perturbations. Data are shown as mean (red line) ± s.e.; data points correspond to individual technical replicates for the indicated perturbation.

(ref. 22). This phosphorylation was believed to occur at threonine 4,539 on the basis of analysis of CDK1 binding motifs. However, the decreased ratio of S4613 to S4618 after treatment with the CDK inhibitor dinaciclib indicated that S4613 (the first serine of the SPYS motif) might be the actual target of phosphorylation by CDK1, or might be cophosphorylated with T4539, given the proximity of these residues in repeat 6.

Plectin is also known to be a substrate of MAP-kinase-interacting serine/threonine-protein kinase 2 (MNK2), which targets a site in repeat 6 distinct from the target site of CDK1 (refs. 22,23). Several of the perturbations we analyzed (selumetinib and PD0325901) target the MEK–ERK pathway, part of the MAPK signaling cascade that could regulate the activity of MNK2. Supporting this, ERK inhibitors have been shown to prevent plectin phosphorylation by MNK2 (ref. 23). In contrast, the mean S4613/S4618 ratio is lower for the p38 MAPK inhibitor losmapimod than for either MEK inhibitor, and p38 MAPK inhibitors do not prevent MNK2 phosphorylation of plectin[23]. Taken together, these considerations suggest that S4613 is phosphorylated by CDK1 and not by MNK2, with the converse being true of S4618.

## Label-free quantification by DIA with Specter is reproducible

DIA allows for the detection of analytes with low relative abundance and does not suffer from the inconsistencies of stochastic precursor selection[24]. We measured an unfractionated *E. coli*

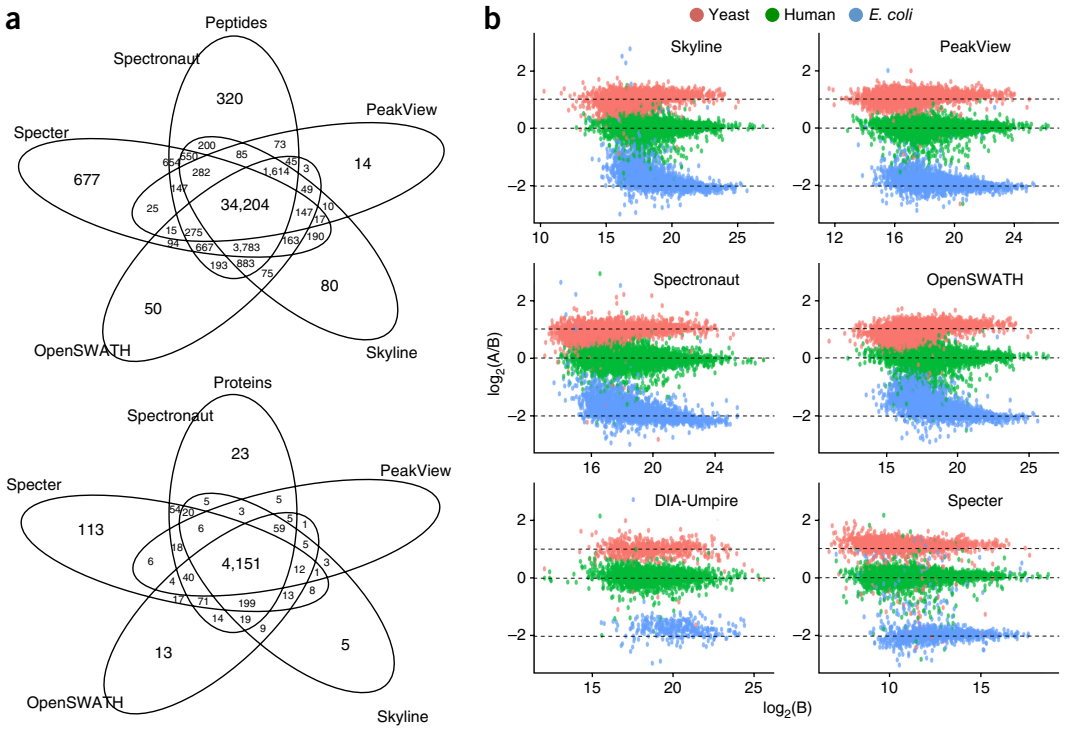


**Figure 6** | Analysis of a public data set by Specter and five other software tools. (**a**) Overlap in peptide and protein identifications from Specter and the four library-based tools. At least two peptides associated with a given protein had to be identified in order for the protein to be considered identified. (**b**) Log ratios of high-precision peptide quantifications (coefficient of variation across replicates < 10%) for the two LFQbench samples from all six tools, plotted against the intensity of sample B. The expected logarithms of these ratios are indicated by the horizontal dashed lines. Specter quantifications shown are the true values, whereas those for all other tools were scaled to a common range by the authors of the LFQbench study.



lysate using both DDA and DIA strategies, each performed back-to-back in triplicate on the same instrument. The DIA runs were analyzed by Specter, using a spectral library containing 48,131 precursors obtained from DDA runs of ten fractions of the lysate (the same library as used in the synthetic spike-in experiment with *E. coli* background described above), whereas the DDA runs were analyzed with MaxQuant.

We found that DIA with analysis by Specter was more reproducible than DDA, owing to the large numbers of precursors identified in one replicate DDA run but not another or with high variability in their quantifications between runs (**Supplementary Fig. 4a**). Identification and quantification in DIA by Specter were highly reproducible, and the total numbers of peptide and protein identifications (12,204 and 1,190, respectively) were comparable to those obtained with DDA (14,407 and 1,350) in the common precursor range of 389–1,015 $m/z$ (we considered a protein as identified only if at least two of its unique peptides were identified). We quantified these observations by Pearson correlation coefficient (**Supplementary Fig. 4b**; average $r^2$ across DDA replicates, 0.72; average $r^2$ across DIA replicates, 0.98). The dynamic range of DDA spanned roughly four orders of magnitude (precursor quantifications ~ $1.6 \times 10^6$ to $2.5 \times 10^{10}$), whereas that of DIA with Specter spanned more than five (~$3.5 \times 10^5$ to $7.4 \times 10^{10}$).

### Comparison to other DIA analysis methods with LFQbench

We used Specter to analyze a publicly available data set generated for LFQbench[9], an R package for the comparison of label-free quantification results of five popular DIA analysis tools: OpenSWATH[3], Skyline[25], Spectronaut[7], DIA-Umpire[4], and PeakView (aka SWATH 2.0). These data were obtained through SWATH[26] analysis of proteomes of three species (human, yeast, and *E. coli*), mixed in two different samples, A and B, at defined ratios, on an AB SCIEX TripleTOF 6600 with 64 variable-width windows.

Using a spectral library provided by the study's authors, Specter identified 40,343 of the 44,294 library peptides, corresponding to 4,733 proteins (we considered a protein as identified only if at least two of its unique peptides were identified in the same sample). The other library-based tools identified 35,517–42,439 peptides and 4,518–4,692 proteins (**Fig. 6a**). DIA-Umpire was not included in this comparison because it is an untargeted method, whereas all others used are targeted and make use of the same spectral library. **Figure 6b** displays the log ratios $\log_2$(A/B) of the most precise peptide quantifications between the two samples (coefficient of variation across replicates < 10%) as reported by the tools as functions of the peptides' intensities in sample B. By following the LFQbench study methods and through comparison with the other tools (**Supplementary Table 4**)[9], we found that Specter had the highest accuracy in quantifying *E. coli* peptides, which have the most extreme expected ratio out of the three species (the first, second, and third tertile mean accuracies of the five non-Specter tools were 0.635, 0.38, and 0.182, versus 0.16, 0.16, and 0.18 for Specter, respectively).

## DISCUSSION

The mixed mass spectra produced by DIA are, in ideal terms, linear combinations of pure spectra. Although the constituents

of such a combination and their abundances are difficult to measure precisely because of shared fragments, biochemical noise, instrument inaccuracy, and inconsistencies in a peptide's fragmentation profile across experiments, we have shown that a linear model enables a principled and effective approach to spectral deconvolution. This allows for the calculation *in silico* of total ion chromatograms for all library precursors without recourse to individual fragment ion traces, which is an important feature in cases where the user's spectral library lacks fragment ion annotations.

Specter can identify and quantify precursors with highly similar library spectra, even when these precursors are coisolated and cofragmented. This has been very challenging for available tools; in fact, several existing methods include rules to explicitly omit such cases from consideration by either discarding shared fragments or excluding all but one member of a given group of precursors with similar spectra[3,5,6].

We illustrated the use of Specter for analysis of the differential regulation of positional phosphoisomers across a series of perturbations of prostate cancer cells by kinase-pathway inhibitors. Other methods for disambiguation of positional isomers of post-translationally modified peptides, such as the IPF algorithm of Rosenberger *et al.*[21], may be advantageous in situations where separate spectra are not available for each isomer, as Specter is a targeted method that requires separate, pre-existing library spectra for each individual analyte one wishes to detect in a given DIA experiment. IPF requires only a single library spectrum representing the fragment ions that are common to all positional isomers of a given peptide sequence. However, this approach does not exploit as much known information as possible and is likely to be less sensitive than Specter in cases where separate spectra are available.

Specter's robustness to incompleteness and noise within a spectral library reduces the need for fractionation and time-consuming curation of fragment ions. Given Specter's use of all features in precursor fragmentation patterns, in general the ideal spectral library for a given DIA experiment should be generated from DDA runs of the samples under consideration with the same instrument.

In the future, we will aim to increase Specter's scope to allow for the characterization of nonlibrary analytes on the basis of correlations between fragment ion elution profiles, in a spirit similar to that of DIA-Umpire[4]. This will expand on the linear model to explicitly account for the linear contributions of unknown analytes to sequential MS2 spectra while simultaneously identifying and quantifying known library members. We also intend to introduce an online interface through which researchers may submit their data for Specter analysis, for users without access to a computing cluster or Apache Spark.

Specter helps to fulfill DIA's promise to provide numbers of peptide and protein identifications similar to those obtained by DDA while offering greater reproducibility across runs and a broader dynamic range. We expect that its high sensitivity and specificity will accelerate the pace of research both into DIA methods generally and into novel applications enabled by the unbiased, reproducible observation and differential quantification of proteins in a broad spectrum of biological contexts.

## METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*


### ACKNOWLEDGMENTS

This work was supported by the NIH (grant U54 HG008097 to J.D.J.). The authors thank S. Tenzer and L. Gillet for providing details of the data sets for the LFQbench study, and K. Clauser for many insightful discussions.


### AUTHOR CONTRIBUTIONS

R.P. conceived of the methodology, created the software, performed the analyses, and wrote the manuscript. S.A.M. conceived and carried out the experiments involving similar synthetic peptides and DIA-DDA comparison, and helped revise the manuscript. A.S.V.J. carried out the experiments and manual analysis of the synthetic phosphopeptide spike-in experiments. J.D.E. and J.G.A. developed the DIA acquisition method. M.J.M., S.A.C., and J.D.J. provided laboratory resources and guidance on the manuscript.

### COMPETING INTERESTS

The authors declare no competing interests.




1. Tabb, D.L. *et al.* Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **9**, 761–776 (2010).
2. Picotti, P. & Aebersold, R. Selected reaction monitoring–based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* **9**, 555–566 (2012).
3. Röst, H.L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32**, 219–223 (2014).
4. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258–264 (2015).
5. Wang, J. *et al.* MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. *Nat. Methods* **12**, 1106–1108 (2015).
6. Bilbao, A. *et al.* Ranking fragment ions based on outlier detection for improved label-free quantification in data-independent acquisition LC-MS/MS. *J. Proteome Res.* **14**, 4581–4593 (2015).
7. Bruderer, R. *et al.* Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics* **14**, 1400–1410 (2015).
8. Bern, M. *et al.* Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Anal. Chem.* **82**, 833–841 (2010).
9. Navarro, P. *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34**, 1130–1136 (2016).
10. Schulz-Trieglaff, O. *et al.* Statistical quality assessment and outlier detection for liquid chromatography-mass spectrometry experiments. *BioData Min.* **2**, 4 (2009).
11. Frewen, B.E., Merrihew, G.E., Wu, C.C., Noble, W.S. & MacCoss, M.J. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.* **78**, 5678–5684 (2006).
12. Ritter, G.L., Lowry, S.R., Isenhour, T.L. & Wilkins, C.L. Factor analysis of the mass spectra of mixtures. *Anal. Chem.* **48**, 591–595 (1976).
13. Likić, V.A. Extraction of pure components from overlapped signals in gas chromatography-mass spectrometry (GC-MS). *BioData Min.* **2**, 6 (2009).
14. Nikolskiy, I., Mahieu, N.G., Chen, Y.-J., Tautenhahn, R. & Patti, G.J. An untargeted metabolomic workflow to improve structural characterization of metabolites. *Anal. Chem.* **85**, 7713–7719 (2013).
15. Du, X. & Zeisel, S.H. Spectral deconvolution for gas chromatography mass spectrometry-based metabolomics: current status and future perspectives. *Comput. Struct. Biotechnol. J.* **4**, e201301013 (2013).

16. Lam, H., Deutsch, E.W. & Aebersold, R. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. *J. Proteome Res.* **9**, 605–610 (2010).
17. Cheng, C.-Y., Tsai, C.-F., Chen, Y.-J., Sung, T.-Y. & Hsu, W.-L. Spectrum-based method to generate good decoy libraries for spectral library searching in peptide identifications. *J. Proteome Res.* **12**, 2305–2310 (2013).
18. Du, X. *et al.* Linear discriminant analysis-based estimation of the false discovery rate for phosphopeptide identifications. *J. Proteome Res.* **7**, 2195–2203 (2008).
19. Shastry, B.S. SNPs in disease gene mapping, medicinal drug development and evolution. *J. Hum. Genet.* **52**, 871–880 (2007).
20. Lawrence, R.T., Searle, B.C., Llovet, A. & Villén, J. Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry. *Nat. Methods* **13**, 431–434 (2016).
21. Rosenberger, G. *et al.* Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. *Nat. Biotechnol.* **35**, 781–788 (2017).
22. Malecz, N., Foisner, R., Stadler, C. & Wiche, G. Identification of plectin as a substrate of p34cdc2 kinase and mapping of a single phosphorylation site. *J. Biol. Chem.* **271**, 8203–8208 (1996).
23. Bouameur, J.-E. *et al.* Phosphorylation of serine 4,642 in the C-terminus of plectin by MNK2 and PKA modulates its interaction with intermediate filaments. *J. Cell Sci.* **126**, 4195–4207 (2013).
24. McQueen, P. *et al.* Whole cell, label free protein quantitation with data independent acquisition: quantitation at the MS2 level. *Proteomics* **15**, 16–24 (2015).
25. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
26. Gillet, L.C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* **11**, O111.016717 (2012).

## ONLINE METHODS

**Constructing the reference spectra library matrix.** For construction of the library spectra matrix $L$, which Specter uses as the design matrix for the underlying linear algebra problem (**Supplementary Note 3**), an instrument mass accuracy parameter $\delta$ is required. For the data analyzed in this paper, all of which were acquired on high-resolution instruments (Thermo Q-Exactive Orbitrap or AB Sciex TripleTOF 6600), $\delta$ was set to 10 p.p.m. for Orbitrap data and 30 p.p.m. for TripleTOF data. Let $S$ be an acquired MS2 spectrum from the DIA run. $S$ is analyzed using only a subset of the provided spectral library, as 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 will be the case if, for example, 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 5 min more or less than the time of the scan (this time window can be omitted or adjusted by the user).

Although 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 its inclusion is highly encouraged in cases where the library and DIA spectra are gathered in similar time frames.

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 $S$ to obtain a vector of intensities whose length equals the number of peaks of $S$ (**Supplementary Note 3**). Each library spectrum is normalized so that its total ion intensity is 1, and these normalized spectra are arranged as the columns of a matrix $L$ whose number of rows equals the length of $S$.

**Finding the optimal combination.** Let $S$ be an MS2 scan from the DIA experiment, represented as a vector of $n$ intensity values. 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 are not present in $S$. Let $L$ be the corresponding matrix of normalized reference spectra constructed above (see also **Supplementary Note 3**). Our aim is to find the non-negative linear combination of the columns of $L$ (the normalized library spectra) that best explains $S$, that is, is closest to it in terms of Euclidean distance. Some peaks of $S$ might not be close to any of the peaks of the reference spectra, as determined using the mass accuracy $\delta$, and these may be discarded from the analysis because they do not affect the determination of the optimal linear combination (**Supplementary Note 4**); that is, the spectrum $S$ is projected to the linear span of the library spectra before analysis.

With these unnecessary peaks removed, the optimal linear combination of the reference spectra is determined as the solution of the corresponding non-negative least-squares problem, which finds the vector $c$ of length $m$ (where $m$ is the number of spectral library members), all of whose entries are non-negative, such that the matrix product of $L$ with $c$ is as close as possible to $S$ in the Euclidean norm among all such non-negative vectors (**Supplementary Note 4**).

**Peptide identification from Specter coefficients.** From the mathematical formulation above, we see that for every MS2 spectrum $S$ acquired in the DIA run, Specter produces a vector $c$ of non-negative coefficients, each of which is associated with a particular precursor in the spectral library. Each coefficient associated by Specter with a library member in a given MS2 spectrum may be directly interpreted as the sum of the intensities of the fragments produced by that member's precursor within that spectrum. This is a straightforward consequence of the fact that the library spectra are normalized to have a total ion intensity of 1 (**Supplementary Note 3**): when such a normalized spectrum $L$ is multiplied by a coefficient $c$ (meaning that the intensities of all of its peaks are multiplied by this constant), the total ion intensity of the resulting scaled spectrum $c \times L$ can be nothing other than $c$. As the aim of Specter is to represent every acquired DIA MS2 spectrum $S$ as a linear combination

$$S = c_1 \times L_1 + c_2 \times L_2 + \ldots + c_m \times L_m + N$$

it follows that the total ion intensity of the $i$th library spectrum $L_i$ in $S$ is simply the Specter coefficient $c_i$. Indeed, the multiplication of a library spectrum $L$ by a coefficient $c$ is the mathematical analog of the physical fragmentation of $c$ molecules of the precursor whose library spectrum is $L$.

When we consider all MS2 spectra sequentially, this gives us an $m \times r$ matrix of coefficients, where $m$ is the number of members of the spectral library and $r$ is the number of MS2 scans. Each row of this matrix is then a time series describing the 'elution profile' of Specter coefficients of a library precursor across the course of the experiment (so that most entries in each row are 0). We consider a library precursor to be identified by Specter if this elution profile contains a peak (local maximum) of at least five consecutive coefficients that are greater than 1 (where coefficients are considered consecutive only if they are calculated relative to sequential MS2 spectra for which the precursor satisfies the conditions as described above). This is a physical constraint that recognizes that total ion intensities less than 1 cannot possibly correspond to meaningful signal.

Because it is possible that peaks such as these may arise by chance, giving rise to false identifications, we used several chromatographic peak scores to rank the quality of our identifications and allow use of a target–decoy approach for estimation of false discovery rates[10]. First, we defined a peak associated with an identified precursor as a local maximum within a consecutive series of at least five Specter coefficients greater than 1; then we defined the largest peak as the peak with the highest summit among all peaks. The four scores associated with the precursor were then (1) the Specter coefficient at the apex of the largest peak, (2) the variance of the coefficients within the largest peak, (3) the skewness of these coefficients, and (4) the kurtosis of these coefficients. Equations for these scores are presented in **Supplementary Note 5**. Taken as a set, these four scores measure the extent to which the largest peak within the precursor's Specter chromatogram resembles an

ideal Gaussian elution profile[10]. Rather than enforcing a strict match to a Gaussian, we used these scores to develop statistics for confident identifications based on the presumably poor peak shapes of false positives. These four scores are combined into a single score via linear discriminant analysis to establish cutoffs to separate target and decoy spectra.

**Peptide quantification from Specter coefficients.** Because the output of Specter is affected by experimental noise and the presence of precursors for which library spectra might not exist, filtering of the Specter elution profile of each precursor is essential to obtain accurate quantifications. To avoid bias arising from parametric filters, we applied a Kolmogorov–Zurbenko filter with three iterations and windows of width three to smooth the Specter elution profile; this is essentially an iterated moving window average[27]. We then calculated the quantification of the precursor as the area under the largest peak of this filtered profile.

**Decoy spectra generation.** We developed our approach to the generation of decoy spectra in the spirit of strategies that generate decoy spectra directly from real library spectra, rather than beginning with random transformations of the sequences of library peptides and subsequently generating decoy spectra based on theoretical fragmentation. To construct the set of decoy spectra used to analyze a given $m/z$ window, we first chose a random subset of all library precursors whose $m/z$ do not fall into this window, where this subset was chosen to have the same size as the set of true library spectra for this window. We then constructed the decoy spectra for this window by shifting the $m/z$ coordinates of all peaks of these non-window spectra by 20 $m/z$. This method combines the main approaches of Lam *et al.*[16] and Cheng *et al.*[17]. To avoid distorting the quantifications of non-decoy library members through the influence of decoy spectra, we used a two-pass approach: first, we used a hybrid target–decoy library to determine linear discriminant score thresholds (based on the set of scores described above and in **Supplementary Note 5**) to achieve a false discovery rate below 1%. We then re-ran Specter with the target library only, and retained only identified library precursors whose linear discriminant scores were above the determined threshold.

**Mass spectrometry data processing.** All raw mass spectrometry data files (in either Thermo RAW or AB Sciex WIFF format) were converted to mzML format in ProteoWizard MSConvert version 3.0.6141 with peak-picking (centroiding). Spectral libraries are accepted in Skyline's blib format[25], which can be constructed from any of the common MS search result file formats[28].

**Python environment and parallelization over MS2 spectra with Apache Spark.** Specter is written in Python 2 and runs on Apache Spark, a highly efficient cluster-computing framework that enables the parallelization of Specter's core algorithm over all MS2 spectra acquired in the course of a DIA run. All analyses in this paper were performed using Python 2.7.11 and Spark 1.6.0 on a computing cluster with 48 identical cores (Intel Xeon CPU E5-2697 v2 at 2.70 GHz) and 250 GB RAM. mzML files are exposed to Python using the run.Reader() function from the Python package pymzml v. 0.7.7. A Python version between 2.7.9 and 3 is required to guarantee compatibility with Spark and the packages used by Specter. A Python list is constructed, each of whose elements is a two-column matrix containing the $m/z$ coordinates and intensities of the peaks of an acquired MS2 spectrum, and this is converted to a Spark resilient distributed data set (RDD) via the parallelize() method. The Spark mapPartitions() method is then applied to this RDD to distribute the analysis of the individual MS2 spectra over the computing cluster. The results are returned to the driver node as a list via the collect() method and subsequently written to a .csv file containing the Specter coefficients of each precursor within each MS2 spectrum for downstream processing. A typical DIA experiment file (~3–5 GB) can be analyzed with a spectral library of ~20,000 precursors in under 30 min by this workflow.

**HEK293T spike-in experiment.** HEK293T cells were cultured in DMEM (Gibco; 11995) supplemented with 10% heat-inactivated FBS (Sigma; F4135). Once cells reached ~95% confluence they were harvested by scraping. Cells were pelleted at 1,000*g* for 2 min. The supernatant was then removed, and the cell pellet was frozen in liquid nitrogen. HEK293T cells were lysed by 5 min of exposure on ice to a lysis buffer (8 M urea, 75 mM NaCl, 50 mM Tris-HCl, pH 8.0, 1 mM EDTA, 2 µg/mL aprotinin (Sigma; A6103), 10 µg/mL leupeptin (Roche; 11017101001), 1 mM PMSF (Sigma; 78830)). The sample was centrifuged for 10 min at 20,000*g*. The protein concentration of HEK293T proteins was determined by BCA assay to be 4.3 µg/µl. 10 mg of protein was reduced (5 mM dithiothreitol, 45 min) and alkylated (10 mM iodoacetamide, 45 min). A Tris-HCl solution (50 mM, pH 8) was used to dilute the samples by a factor of 4 to reach a concentration of 2 M urea. A two-step digestion protocol was used to digest the lysate: Lys-C was used in a 1:50 enzyme-to-substrate ratio (Wako Chemicals; 129-02541) for 2 h at 30 °C, then the lysate was digested overnight at room temperature with trypsin in a 1:50 enzyme-to-substrate ratio (Promega; V511X) on a shaker. Formic acid (FA; 0.5% final concentration) was added to stop the digestion. The sample was split into four aliquots and loaded onto four 100-mg-capacity C18 Sep-Pak cartridges (Waters) for desalting. The four aliquots were eluted with 50% acetonitrile (ACN)/0.1% FA, pooled together, and vacuum-concentrated to dryness. The HEK293T digest was resuspended using 0.1% TFA, and a mixture containing 96 synthetic peptides at known individual concentrations (**Supplementary Table 5**) was spiked into it to generate a five-point calibration curve. Each point was designed to contain 1 µg of HEK293T digest and 6.75, 13.5, 27, 54, or 108 ng of total peptide on column.

All samples were analyzed with an Orbitrap Q-Exactive HF Plus (Thermo Fisher Scientific) mass spectrometer coupled to a nanoflow Proxeon EASY-nLC 1000 UHPLC system (Thermo Fisher Scientific). The mass spectrometer was used in positive mode and was equipped with a nanoflow ionization source (James A. Hill Instrument Services, Arlington, MA); the spray voltage was set at 2.00 kV. The LC system, the column, and the electrospray voltage source (platinum wire) were connected via a stainless steel cross (360 µm; IDEX Health & Science; UH-906x). The column was heated to 50 °C. A volume of 3 µl was injected onto an in-house packed 20 cm × 75 µm diameter C18 silica picofrit capillary column (1.9-µm ReproSil-Pur C18-AQ beads, Dr. Maisch GmbH, r119.aq; Picofrit 10-µm tip opening, New Objective, PF360-75-10-N-5). The mobile phase had a flow rate of




doi:10.1038/nmeth.4643

250 nL/min and consisted of 3% ACN/0.1% FA (solvent A) and 90% ACN/0.1% FA (solvent B). The column was conditioned before each sample injection. Peptides were separated using the following LC gradient: 0–3% B in 3 min, 5–40% B in 50 min, 40–90% B in 1 min, stay at 90% B for 5.5 min, and 90–50% B in 30 s. DDA and DIA data were acquired on the same instrument. For the MS1 scans the resolution was set at 60,000 at 200 *m*/*z* and the automatic gain control (AGC) target was $3 \times 10^6$ with a maximum inject fill time of 20 ms. For DDA, MS2 scans on the top 12 peaks doubly charged and above were acquired at a resolution of 15,000, AGC target of $5 \times 10^4$ with maximum inject fill time of 50 ms. Isolation widths were set to 1.5 *m*/*z* with a 0.3 *m*/*z* offset. The normalized collision energy (NCE) was set to 27 and dynamic exclusion was set to 10 s. For DIA, an overlap DIA method was used with 56 × 22 *m*/*z* isolation windows covering the 400–1,000 *m*/*z* range. In this method the isolation windows in two consecutive cycles have an offset of 11 *m*/*z*. The default charge state was 4, the resolution was 30,000 at 200 *m*/*z*, the AGC target was $1 \times 10^6$, the maximum inject fill time was 50 ms, the loop count was 27 and the NCE was set to 27.

We generated a spectral library by first searching the DDA runs with Spectrum Mill v. B.06.01.201 using a FASTA containing the 96 synthetic peptide sequences and the UniProt human protein sequences (version dated 17 October 2014). Results were auto-validated to a false discovery rate of 1%. This yielded a PepXML search result file, which was loaded into Skyline v. 3.6.0.10162 to generate a spectral library of 29,248 HEK293T and 85 synthetic precursors (11 of the 96 were not identified) in blib format[29].

Whereas manual quantifications are determined from the extracted ion chromatograms of only preselected fragment ions, quantification by Specter incorporates every peak present in its library spectrum. Thus, potentially significant differences between the quantifications are to be expected, so the similarity between the two modes of analysis is more appropriately measured by the correlation between quantifications for each peptide (**Fig. 2c**) than by the absolute differences between the quantifications.

**False discovery rate estimation.** Of the 29,333 HEK293T and synthetic library precursors, 8,867 were cumulatively identified by Specter in all three replicate DIA runs, whereas 297 *E. coli* peptides were cumulatively identified. Thus, if a decoy library of the same size as the yeast/bovine library were constructed through selection of 29,333 of the 48,131 *E. coli* library spectra uniformly at random, the estimated false discovery rate would be

$$(297/48{,}131) \times 29{,}333/(8{,}867 + (297/48{,}131) \times 29{,}333) \approx 0.02$$

**Synthetic peptide experiment.** Synthetic peptides were obtained through New England Peptide Inc. (Gardner, MA), predissolved in 30% ACN/0.1% FA, and diluted to 100 µM in the same solvent.

To obtain library spectra for the synthetic peptides, we injected each peptide individually at 50 fmol, 200 fmol, and 1 pmol on column with wash runs in between peptides. On-line liquid chromatography was performed with an EASY nano-LC 1000 UHPLC (Thermo). Separation was performed on a 20-cm, 75-µm inner-diameter column, packed in-house with 1.9-µm C18-AQ beads (Dr. Maisch) with a gradient from 2% ACN to 55% ACN over 20 min. The data were acquired on a Q-Exactive Plus mass spectrometer (Thermo Fisher Scientific) in data-dependent top 12 mode using a resolution of 70,000 for MS1 and 17,500 for the MS2 scans. Dynamic exclusion was disabled to obtain MS2 multiple times for each precursor across the peak. The resulting raw files were searched with Spectrum Mill v. B.06.01.201 with a FASTA containing only the sequences of the synthetic peptides and common contaminants. The best-scoring spectrum for each precursor was then chosen to serve as the precursor's library spectrum.

DH5α *E. coli* were grown in Luria broth at 37 °C overnight. Cells were pelleted by centrifugation, washed once with cold PBS, flash-frozen in liquid nitrogen, and stored at –80 °C until processing. For generation of the *E. coli* lysate digest, the cell pellet was thawed on ice. Lysozyme (Sigma) was added to the thawed pellet, and the mixture was placed on ice with periodic vortexing until viscous. The cells were resuspended in 8 M urea, 50 mM ammonium bicarbonate plus protease inhibitors (Roche), and the solution was sonicated with a probe sonicator for 2 min, 3 s on, 2 s off, until no longer viscous. After centrifugation at 15,000*g* for 30 min at 4 °C, protein concentration was measured by Bradford assay (Bio-Rad). Disulfide bridges were reduced with 10 mM TCEP (tris(2-carboxyethyl)phosphine; Thermo) and alkylated with 10 mM iodoacetamide (Thermo) for 30 min at room temperature in the dark. The lysate was diluted to 1.5 M urea with 50 mM ammonium bicarbonate and digested overnight with a trypsin-to-substrate ratio of 1:100. The digest was desalted on C18 Sep-Pak cartridges (Waters). After vacuum centrifugation, dried peptides were resuspended to 1 mg/mL in 30% ACN/0.1% FA and stored at –80 °C.

To generate library spectra of the *E. coli* digest, we fractionated peptides using a StageTip[30] packed with sulfonated poly-(styrene-divinylbenzene) resin (SDB-RPS; EMPORE). 100 µg of digest was fractionated starting from 20 mM ammonium hydroxide, pH 10, with the percentage of can increased by steps of 5, 10, 12.5, 15, 17.5, 20, 25, 30, 35, and 50%. Assuming equal mass distribution, 1 µg of fractionated digest was analyzed by LC-MS2. Data were acquired on the same instruments as listed above with changes to the LC gradient and the data acquisition. Peptides were separated on-line with an 85-min gradient from 6% 0.1% FA (buffer A) to 30% 0.1% FA/90% ACN (buffer B), followed by an increase to 60% buffer B over 9 min. The mass spectrometer was set to acquire data at a resolution of 70,000 and an AGC setting of $3 \times 10^6$ for MS1. MS2 resolution was 17.5K, AGC $5 \times 10^4$, and maximum inject time of 100 ms. The top 12 ions identified within the precursor scan of 300–2,000 *m*/*z* that were at least doubly charged were selected for high-energy collisional dissociation at an NCE of 25. Raw files were searched by SpectrumMill v. B.06.01.201 using the NCBI *E. coli* K12 DH10B FASTA sequence database (a FASTA for DH5α was unavailable) and auto-validated to a false discovery rate of 1%. The results were exported as a PepXML summary file, which was imported into Skyline v. 3.6.0.10162 to generate a spectral library consisting of 48,131 precursors. This spectral library was exported in blib format[25,28] for later use by Specter.

For spike-in experiments, synthetic peptide mixtures were constructed according to **Supplementary Table 3**. We created 20 injections worth of peptide mixtures for each of the zero, single, or double drop-out samples. The pools were then equally

divided either as synthetic peptides alone or into the same *E. coli* digest described above. In either case, roughly equal amounts of synthetic peptide material were loaded on column, regardless of whether *E. coli* background was present. For runs containing lysate background, approximately 1.5 µg of *E. coli* digest was loaded on column. DDA was performed on a Q-Exactive Plus HF mass spectrometer, where MS1 scans were measured at a resolution of 60,000 and an AGC setting of $3 \times 10^6$ and maximum injection time of 20 ms. MS2 scans on the top 15 peaks doubly charged and above were acquired at a resolution of 15,000, AGC target of $5 \times 10^4$ and maximum inject time of 50 ms. Isolation widths were set to 1.7 Th with a 0.3-Th offset. NCE was set to 28 and dynamic exclusion was set to 15 s.

DIA data were acquired with MS1 parameters as above (range: 300–1,200 *m*/*z*) and then using 22-Th-wide windows for MS2 with a default charge state of 4 at a resolution of 30,000. The AGC target was set as $1 \times 10^6$, maximum inject time was set to 50 ms, and a loop count of 27 was used. NCE was set to 27. A total of 56 × 22 Th DIA windows were used to traverse the *m*/*z* range of 400–1,000, with the range actually traversed twice but the windows offset by 50%. The window centers can be found in **Supplementary Table 1**. LC and nanospray parameters were identical to those described by Abelin *et al.*[31].

**Positional isomers in drug-perturbed cellular systems.** Sample preparation and experimental procedures were identical to those described by Abelin *et al.*[31], and DIA runs were performed with the same parameters used in the synthetic peptide experiment described above. Drug treatments and concentrations are shown in **Supplementary Table 6**. Ten randomly chosen phosphoenriched samples of PC3 cells treated with the perturbations highlighted in bold in **Supplementary Table 6** were measured by DDA (acquired with the same settings as the DDA runs described above). Results were searched using Spectrum Mill as above with a FASTA containing the 2014 UniProt Human proteome and 150 common contaminants. Phosphorylations of serine, threonine, and tyrosine were set as variable modifications. We imported the resulting pepXML files into Skyline to construct a redundant blib containing multiple peptide–spectrum matches for each precursor. Search results were further filtered by variable modification location score so that only spectra for which phosphosites could be unambiguously localized were retained, and the highest-scoring spectra for each precursor were extracted from the redundant blib to produce a nonredundant spectral library consisting of only confidently localized phosphopeptides. Specter was applied to the 84 DIA runs using this spectral library and a 10 p.p.m. mass accuracy.

***E. coli* DDA analysis with MaxQuant.** RAW files obtained from triplicate DDA runs of the unfractionated *E. coli* digest described above were imported into MaxQuant v. 1.5.5.1. These were searched using the same FASTA as used by Spectrum Mill above with all default settings for Orbitrap instruments ("Match between runs" was disabled in order to assess replicate reproducibility). Results were analyzed using the evidence.txt output table. Only precursors with posterior error probability < 0.01 were retained, and the top-scoring MS2 spectrum for each of these precursors within each replicate was used to determine precursor quantifications within each run. Protein identifications (based on all three replicates) were determined from the proteinGroups.txt output table.

**LFQbench data set.** All data were downloaded from ProteomeXchange (data set PXD002952). Raw WIFF files from the HYE124 data set (with 64 variable-width windows on a Triple TOF 6600) were converted to mzML with ProteoWizard as described above. We used a spectral library provided by the study's authors (ecolihumanyeast_concat_mayu_IRR_cons_openswath_64w_var_curated.csv) that consisted of precursors with annotated fragment ions in CSV format compatible with OpenSWATH. The mass accuracy parameter δ was set to 30 p.p.m.

For comparisons to other analyses, only the results from the first iteration of the LFQbench study were used. This was based on the consideration of the optimizations and open discussion among software developers that took place for the second iteration, in which we did not participate.

**Code availability.** Specter is available as an open-source software tool at https://github.com/rpeckner-broad/Specter.

**Life Sciences Reporting Summary.** Further information on experimental design is available in the **Life Sciences Reporting Summary**.

**Data availability.** All original mass spectrometry proteomics data used to support the conclusions of this study have been deposited to the ProteomeXchange Consortium via the PRIDE[32] partner repository with the data set identifier PXD006722. The data used from the LFQbench study are available from the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD002952. Source data for **Figures 2–6** are available online.


27. Yang, W. & Zurbenko, I. Kolmogorov-Zurbenko filters. *Wiley Interdiscip. Rev. Comput. Stat.* **2**, 340–351 (2010).
28. Deutsch, E.W. File formats commonly used in mass spectrometry proteomics. *Mol. Cell. Proteomics* **11**, 1612–1621 (2012).
29. Frewen, B. & MacCoss, M.J. Using BiblioSpec for creating and searching tandem MS peptide libraries. *Curr. Protoc. Bioinformatics* **20**, 13.7.1–13.7.12 (2007).
30. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
31. Abelin, J.G. *et al.* Reduced-representation phosphosignatures measured by quantitative targeted MS capture cellular states and enable large-scale comparison of drug-induced phenotypes. *Mol. Cell. Proteomics* **15**, 1622–1641 (2016).
32. Vizcaíno, J.A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, D447–D456 (2016).






doi:10.1038/nmeth.4643

nature research

Corresponding author(s): Jacob D. Jaffe

# Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

## Experimental design

1. Sample size

Describe how sample size was determined.

As our study presents a new method for analyzing mass spectrometry data, rather than an attempt to illustrate a statistically significant biological phenomenon , no sample size considerations were relevant.

2. Data exclusions

Describe any data exclusions.

We excluded data that would be considered "noise" in mass spectrometry data, i.e. signals of very low intensity relative to the main signal of interest. This was based on pre-established quantitative criteria.

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

All mass spectrometry experiments were performed at least in duplicate (most often in triplicate), with high reproducibility in almost all cases.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

One of our experiments did involve assignments of synthetic peptides to groups, which were randomly chosen by co-author Samuel Myers.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

For the synthetic peptide experiment, group assignment was chosen randomly by co-authors S.M. while analysis was performed by R.P., who had no knowledge of the members of each group.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

| n/a | Confirmed |
|---|---|
| ☐ | ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.) |
| ☐ | ☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ A statement indicating how many times each experiment was replicated |
| ☒ | ☐ The statistical test(s) used and whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ A description of any assumptions or corrections, such as an adjustment for multiple comparisons |
| ☒ | ☐ Test values indicating whether an effect is present<br>*Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.* |
| ☒ | ☐ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range) |
| ☐ | ☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation) |

*See the web collection on statistics for biologists for further resources and guidance.*





Nature Methods: doi:10.1038/nmeth.4643

## Software

Policy information about availability of computer code

7. Software

| | |
|---|---|
| Describe the software used to analyze the data in this study. | The purpose of our study was to present a new software tool. As stated in the manuscript, our custom code is available on GitHub at github.com/rpeckner- broad/Specter. |

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

## Materials and reagents

Policy information about availability of materials

8. Materials availability

| | |
|---|---|
| Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party. | No unique materials were used. |

9. Antibodies

| | |
|---|---|
| Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species). | No antibodies were used. |

10. Eukaryotic cell lines

| | |
|---|---|
| a. State the source of each eukaryotic cell line used. | PC3 cell lines were obtained from ATCC (catalog number CRL-1435). |
| b. Describe the method of cell line authentication used. | DNA fingerprinting was applied using the following SNPs: Chromosome SNP ID Position<br>5 rs10037858 156766708<br>1 rs532905 189620278<br>1 rs2229857 154573967 20 rs6104310 44354538 18 rs9304229 38247872 6 rs2273827 160211339 5 rs2036902 49696932<br>1 rs6679393 31770944<br>5 rs2369754 99156362<br>1 rs1052053 156202173 2 rs6726639 112753097 11 rs2512276 124115370 17 rs6565604 79589242 7 rs6972020 68182022<br>8 rs13269287 124453662 1 rs10888734 52266242 7 rs6966770 115895718 7 rs2639 6066461<br>2 rs10186291 112748514 2 rs7598922 39082344<br>4 rs2709828 152355268 2 rs1131171 232326417 4 rs7664169 99037859 17 rs1437808 4175846<br>3 rs11917105 183371250 12 rs10876820 55978465 5 rs2910006 140590766 24 AMG_3b 6737949<br>14 rs8015958 67086676 13 rs3105047 55937194 3 rs5009801 101058775 6 rs9277471 33053682 18 rs3744877 77894844 9 rs1549314 127910307 6 rs9369842 48994615 5 rs390299 153363334 2 rs1734422 10932207 22 rs9466 38273749<br>19 rs4517902 29851078 13 rs6563098 79887237 9 rs965897 77175017 |
| c. Report whether the cell lines were tested for mycoplasma contamination. | The cell lines were tested for mycoplasma contamination. |
| d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use. | None of the cell lines we used are listed in the ICLAC database. |

## Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

| | |
|---|---|
| Provide all relevant details on animals and/or animal-derived materials used in the study. | No animals were used. |

Policy information about studies involving human research participants

12. Description of human research participants

| | |
|---|---|
| Describe the covariate-relevant population characteristics of the human research participants. | The study did not involve human research participants. |





Nature Methods: doi:10.1038/nmeth.4643