Quantitative studies of protein abundance rarely span more than a small number of experimental conditions and replicates. In contrast, quantitative studies of transcript abundance often span hundreds of experimental conditions and replicates. This situation exists, in part, because extracting quantitative data from large proteomics datasets is significantly more difficult than reading quantitative data from a gene expression microarray. To address this problem, we introduce two algorithmic advances in the processing of quantitative proteomics data. First, we use space-partitioning data structures to handle the large size of these datasets. Second, we introduce techniques that combine graph-theoretic algorithms with space-partitioning data structures to collect relative protein abundance data across hundreds of experimental conditions and replicates. We validate these algorithmic techniques by analyzing several datasets and computing both internal and external measures of quantification accuracy. We demonstrate the scalability of these techniques by applying them to a large dataset that comprises a total of 472 experimental conditions and replicates.

Computational Challenges. The first computational challenge entails finding sets of data points that correspond to quantitative measurements of specific peptides. These patterns are known as extracted ion chromatograms (XICs) (4). Each XIC is a series of peaks that occur in a narrow m/z range over an interval of time. A typical scan contains many thousands of XICs, each corresponding to a different peptide from a complex mixture of proteins, along with additional data points due to various sources of noise. The area under an XIC measures the relative abundance of the corresponding tryptic peptide (see Fig. 1).

The second computational challenge entails determining the identity of the peptide associated with an XIC. Specifically, if one of the peaks in this series has a fragmentation spectrum, the spectrum can be used to search a database of predicted spectra based on the genome sequence of the species under study (5–7). From a high-scoring match to the database, one can determine the sequence of the tryptic fragment and the protein from which the fragment originated. We note that such database search is in itself a difficult computational challenge, one that is often addressed independently of quantification. We rely on existing algorithms for database search (5, 6).

The third computational challenge in quantification entails handling hundreds of LC-MS/MS scans that correspond to replicates and experimental conditions. To obtain relative protein abundance across the scans, XICs corresponding to the same peptides must be identified in each scan. Nonlinear variation in retention time of these XICs due to slight differences in chromatography, along with measurement error in m/z, complicates this process. A solution requires a multiscan alignment step to correct for these sources of variation. As the number of scans increases, it becomes significantly more difficult to assure correctness of the global alignment (8).

Previous Work. Existing methods do not adequately address the computational challenges of XIC detection and multiscan alignment and are limited to a small number of scans. Several of these
methods rely on image processing to collect quantitative data from LC-MS/MS scans (9–12). Because image processing typically requires computation on all individual pixels, these methods require a large amount of computation per LC-MS/MS scan. Furthermore, they require expensive image-correlation computations to nonlinearly align LC-MS/MS scans (13). Our algorithmic techniques avoid pixel-level image representations (see below).

To address the limitations of image-based methods, more recent approaches have relied on complex algorithms that process individual spectra one retention time point at a time and combine those results (14–16). We skip this individual scan-processing step, representing the data in two dimensions: retention time and m/z. Our algorithm is much simpler because it exploits structure in the data across both dimensions. Furthermore, existing methods rely on iterative optimization techniques to align scans. These iterative techniques become increasingly less reliable and computationally more expensive as problem dimensionality increases with the introduction of additional scans (8). In this work, we present a technique that is not susceptible to these dimensionality problems and demonstrably scales to hundreds of LC-MS/MS scans.

Results

Validation by Using Spiked-In Proteins. We tested whether our algorithm could accurately track the known abundance of several proteins across multiple LC-MS/MS scans by analyzing the dataset generated by Mueller et al. (16). The bulk human serum background was sufficiently complex to provide a challenge for our algorithm. Specifically, we observed 18,702 XIC groups that occurred in at least two of the three replicates of each dilution in three or more of the six dilution datasets. The large number of grouped XICs indicated that many different tryptic fragments were present, requiring our algorithm to accurately discern XICs from the spiked-in proteins. Next, we correlated the relative protein-abundance data computed by our algorithm from the LC-MS/MS data with the known concentrations of the spiked-in proteins. We observed nearly perfect correlations, with $R^2$ values between 0.97 and 0.99 (see Fig. 2). These results indicate accurate quantification of the relative abundance of specific proteins in a complex mixture. The entire analysis required only 2 min and 4 s of time on a CPU with 4 cores, compared with the 105 min of time reported by Mueller et al. (16) for their method on a CPU with a single core. Accounting for differences in computer configurations, we estimate a processing time improvement of >12-fold.

Validation on a Large Dataset. To test whether our algorithm could quantify large numbers of proteins across hundreds of experimental conditions, we analyzed the dataset generated by Foss et al. (8). This dataset comprises 472 LC-MS/MS scans of total yeast protein. Four hundred twenty eight of the scans represent two biological and two technical replicates for each of 107 progeny from a cross between two yeast strains.

Our algorithm detected 14,208 XICs that grouped across 40 or more progeny. 11,888 of these XIC groups had at least one associated MS/MS spectrum. For 3,692 XIC groups, a peptide was identified by database search (see Methods). These peptides represented 635 unique proteins. The ability to quantify 635 proteins across 40 or more progeny represents a $\approx$3-fold improvement in proteome coverage over the previous analysis of this dataset (see supporting information (SI) Fig. S1).

We next examined the quality of protein quantification. First, we observed a median correlation of 0.98 between technical replicates and 0.95 between biological replicates (see Fig. S2), showing that quantification was highly reproducible. Second, for nine proteins previously measured in the parent strains by quantitative Western blot analysis, the relative abundance was in agreement with the results computed by our algorithm from the LC-MS/MS data (see Fig. S2). Third, we took advantage of genotyping data for the progeny strains to carry out linkage analysis of protein abundance for the 635 proteins measured in at least 40 strains. Linkage analysis tests for significant differences in abundance between groups of strains divided by genotype at a given genetic marker. Errors in XIC detection and grouping would result in poor quantitative profiling of protein
protein abundance by both algorithms. We observed this high agreement with the expected fold differences in a complex dataset. This correlation exceeded the Spearman’s correlation of $r^2 > 0.69$ between our protein abundance ratios and those reported by Baek et al. (see Fig. S3).

Fig. 2. Comparison of known protein concentration with relative abundance measured by mass spectrometry. Measured protein abundance is plotted on a log–log scale against known femtomole concentration for each of six nonhuman proteins spiked in to human serum by Mueller et al. The lines show best fit by regression, and the corresponding log–log scale slopes and correlation values ($R^2$) for both the log–log scale and the original scale are shown below each plot.

Isotope-Labeled Data Validation. To test whether our algorithm could also quantify isotope-labeled data, in addition to the label-free data examined above, we analyzed an isotope-labeled dataset by Baek et al. (17). This dataset contained a complex mixture of labeled and unlabeled peptides. Our algorithm detected 119,140 paired XICs for which the light or the heavy XIC could also quantify isotope-labeled data, in addition to the label-free data examined above, we analyzed an isotope-labeled dataset by Baek et al. (17). This dataset contained a complex mixture of labeled and unlabeled peptides. Our algorithm detected 119,140 paired XICs for which the light or the heavy XIC.
Discussion

We have devised and implemented algorithms for the analysis of large-scale proteomics data that differ from previous work in their ability to scale to hundreds of LC-MS/MS scans. The algorithm we present scales to very large datasets because of its reliance on a space-partitioning data structure that accelerates planar orthogonal range queries. Space-partitioning data structures and planar orthogonal range queries are well-studied ideas in computer science. They are of tremendous practical utility and are ubiquitous in computational geometry, computer graphics, and geographic information systems. Furthermore, there exists a significant body of work on practical aspects of their implementation, engineering, and use (19). Graph theoretic approaches, machine learning, signal processing, and statistics have made significant inroads into proteomics, and we expect the same of computational geometry techniques.

We note that the techniques presented here are not limited to measurements of protein-abundance data and posttranslational modifications, but they also apply to quantitative analysis of other biomolecules with mass spectrometry. We expect that they can be readily adapted to quantitative measurement of metabolite abundance and lipid abundance. Adapting these techniques to these other types of molecules will be the subject of future work. Combined with advances in experimental mass spectrometry techniques, we expect that the algorithms developed here will play a key role in providing a more complete picture of cellular physiology, thereby enabling sensitive and accurate identification of genes and pathways involved in disease.

Methods

Data Structure. Central to all of the algorithms below is a data structure that indexes points bounded on two of their dimensions by time and m/z. The data structure must support the following interface: RangeQuery(D, t1, t2, m1, m2) uses the data structure D to conduct a planar orthogonal range query, returning objects (t, m/z) in a rectangular region defined by t1 < t < t2 and m1 < m/z < m2.

In computer science, there are many data structures that support this interface, each with its own space, speed, simplicity, and efficiency tradeoffs. Examples include kd-trees, range trees, quad-trees, binary space partitioning interfaces, each with its own space, speed, simplicity, and efficiency tradeoffs. For this work, we use a kd-tree because it occupies linear space with respect to the number of data points and requires O(\sqrt{N}) time to conduct a single range query, where N is the number of (t,m/z) objects (millions in a typical scan).

Ion Chromatogram Extraction. The first step of processing an LC-MS/MS scan is to remove peaks that are caused by noise (see Fig. S4A). Given a kd-tree on all of the data points, the algorithm iterates through each peak and performs a planar orthogonal range query with specified width in retention time and height in m/z around the peak. If the number of peaks returned by the query exceeds a threshold, and the peak is above a nominal absolute intensity threshold, the peak is labeled as signal. Otherwise, the peak is labeled as noise. For the datasets analyzed here, we selected an absolute intensity threshold of 100, which we observed by visual inspection to remove only low-intensity noise peaks. After this process, the peaks labeled as noise are removed from further processing.

To find XICs, the algorithm uses planar orthogonal range queries and an undirected graph structure. First, the algorithm makes each peak a node in the graph. Second, the algorithm constructs a kd-tree on all of the signal points. Next, the algorithm iterates through each signal peak and connects the current node to any signal peak nodes returned in a query with specified width in retention time dimension and height in m/z. Last, the algorithm finds XICs by computing the connected components of the constructed graph. An individual connected component corresponds to an XIC (see Fig. S4B).

An XIC has a start retention time and an end retention time. Furthermore, an XIC may have one or more peaks with fragmentation spectra. These can be used to determine the sequence of the corresponding tryptic peptide. An XIC also has an m/z value that corresponds to the average m/z value of all of the peaks grouped into the XIC, and a retention time that corresponds to the retention time of the most-intense peak in the XIC. Because XICs have a center point, we also index them by using a kd-tree data structure.

Extracted Ion Chromatogram Alignment and Grouping. Before grouping the XICs in multiple LC-MS/MS scans, the algorithm first aligns each scan to a reference scan via simple translation. This alignment compensates for any differences in when the samples began to elute out of the LC column. The algorithm iterates through each XIC in the current scan and finds the nearest XIC in the reference scan that is within a rectangular window with specified width in retention time and height in m/z. For this nearest XIC in the reference scan, the algorithm also computes the reciprocal nearest XIC in the current scan. If this reciprocal XIC is the same as the current XIC, the difference in retention time between the current XIC and the nearest reference XIC is stored in a list. After all of the XICs in the current scan are processed, the median of these differences in retention time is used to translate each XIC in the current scan to the reference search.

Once each scan is aligned to the reference scan via translation, XICs are grouped across these scans. First, each XIC belonging to a scan is labeled by using the LC-MS/MS scan’s identifier. Then, each XIC from that scan is combined into a larger set of labeled XICs from all scans (see Fig. S5A). Last, we use the same technique applied to find XICs from peaks to group XICs across scans. Each XIC starts out as a node in a graph. Planar orthogonal range queries between the start and end of each XIC in retention time are used to connect XICs in a graph. Connected components in this graph correspond to grouped XICs (see Fig. S5B). The planar orthogonal range queries used in this step automatically compensate for any nonlinear differences in the positions of XICs. Furthermore, the range queries can be expanded far beyond the start and end of the XICs to compensate for more-severe differences in XIC retention time.

Collecting Relative Abundance Data. Once these XICs are grouped across scans, relative abundance values are computed, and an amino acid sequence and protein identity is assigned to the group. Relative abundance values across scans are computed as the areas under each XIC in the group. These relative protein abundance values are normalized by using median of medians normalization to adjust for differences in overall scan intensity (21). Then, the entire XIC group is assigned an amino acid sequence for a particular tryptic peptide by database search. Database search is conducted by using all or a subset of fragmentation spectra within an XIC group.

Once the XIC group is assigned an amino acid sequence and protein identity, relative abundance data are available for a known tryptic peptide from a known protein. Because individual proteins are digested into several tryptic fragments, there may be many different XIC groups for a single protein, but in most applications, only the relative abundance of a protein is required. Because each tryptic fragment from a protein ionizes with varying efficiency, the quantitative values in each XIC group cannot be combined by simple averaging. Instead, a representative XIC group is selected for each protein. Specifically, we select the XIC group with highest signal-to-noise ratio. We note that other criteria, tailored to specific applications, can be used to select representative XIC groups.

Stable Isotope-Labeled Data Processing. An alternative to label-free quantification is isotope-labeled quantification. In this experimental technique, one of two experimental samples is labeled with a heavier isotope tag (22). Because both the unlabeled and isotope-labeled tryptic fragments are subject to the same chromatographic conditions, they elute out of a liquid chromatography column at approximately the same retention time. Isotope-labeled quantification reduces to finding pairs of XICs within a single scan at the same retention time that are spaced according to the charge of the tryptic fragment and the weight difference of the heavier isotope tag. The ratio of the areas of these XICs measures the relative abundance of the corresponding tryptic fragment between the experimental samples.

Instead of aligning and grouping XICs across scans, the algorithm handles isotope-labeled data by grouping XICs within a scan. This is accomplished by reciprocal planar orthogonal range queries between XICs. XICs are processed in increasing m/z order, starting with XICs corresponding to the lighter variant of the peptide. XICs that have already been paired are skipped. Based on the charge of the XIC, the expected isotopic spacing of the heavier variant is computed. An orthogonal range query between the start and end of the current XIC and an m/z width creates a putative paired XIC. The longest of the returned XICs is used to conduct a reciprocal planar orthogonal range query between the start and end of the putative paired XIC. If the current XIC is returned by this reciprocal query, the two XICs are paired.

Fragmentation spectra assigned to the paired XICs are used to determine the amino acid sequence and protein identity associated with the pair. Be-
cause proteins are digested into several tryptic fragments, a dataset may contain several paired XICs per protein. Unlike label-free data, the abundance ratios computed by these pairs are comparable. The median ratio from all of these pairs can be used to compute a robust estimate of the abundance ratio for an individual protein.

Datasets. We used four datasets to test and validate our algorithms. The first dataset is a spike-in dataset used in a previous study by Mueller et al. (16). The dataset contains six nonhuman proteins added at six different known concentrations to a background sample of bulk human serum (see Table S1). Three replicates of each dilution were collected by using an FT-LTQ ThermoElectron mass spectrometer and an Agilent 1100 chromatographic separation system. In total, this dataset consists of 18 LC-MS/MS scans and is 15 GB in size. We downloaded the data from http://prottools.ethz.ch/muellelu/web/LatinSquare.Data.php.

The second dataset, described by Foss et al. (8), measured total unfraccionated cellular proteins from 107 genotyped segregants from a cross between two parental strains of yeast (BY4716 and RM11–1a). Four replicate LC-MS/MS scans were carried out for each segregant. The data also include 10 replicates of each parent strain and 2 replicates of each of six gas-phase fractions from each parent strain. In total, this dataset includes 472 LC-MS/MS scans and is 42 GB in size. The dataset was generated by using a ThermoElectron Corp. LTQ-FT mass spectrometer and a Michrom Bioresources Paradigm MS4B MDLC nanoflow liquid chromatography system. We obtained the data from the authors.

The third dataset, used to test performance on isolate-labeled data, derives from a study by Baek et al. (17) on the effect of microRNAs on protein levels. In this dataset, the unlabeled sample originated from the cytosolic fraction of mouse cultured cells. The isotope-labeled sample originated from the cytosolic fraction of mouse cultured cells with a miR-223 knockout. The dataset was generated with a ThermoElectron LTQ-OrbitrapXL and consists of 16 LC-MS/MS scans from 16 gel fractions. The sample was labeled with Arg-6 and Lys-6. We obtained the data from the authors.

The fourth dataset used a deuterated isotope label (5-Da shift) (23). To generate these data, samples of histone H3 were extracted from MEF and mESC. The MEF sample was derivatized with propionic anhydride (D0, 56 Da), whereas the mESC sample was derivatized with an isotopically labeled propionic anhydride (DS, 61 Da). The data were collected with a ThermoElectron LTQ-Orbitrap and consisted of single LC-MS/MS scan. This dataset was generated as part of this study.

Protein Identification. Amino acid sequence and protein identity were assigned by using the X! Tandem database search algorithm (version 08-02-01-3 (6), e-value <0.1). For the Foss et al. data, we additionally used the OMMSA algorithm (version 2.1.1 (5), e-value <0.1).

CPU and Availability. Algorithm timing and peak memory usage was collected on a machine with two dual-core AMD Opteron 2220 2.8 GHz Processors and 32 GB of RAM (four CPU cores total). All of the algorithms described in these methods were implemented in the Princeton LC-MS/MS Data Viewer. Complete source code is available at http://compbio.cs.princeton.edu/pview.

Supporting Information

Khan et al. 10.1073/pnas.0904100106

SI Text

To further illustrate the simplicity of the algorithms presented in the text, we include pseudocode for each algorithm below. We assume an implementor has a well-designed kd-tree implementation and an implementation of a connected-components algorithm. Note that mass spectrometers typically operate in one of three modes: raw profile, profile, and centroid. In raw profile mode, the signal in a scan is finely sampled at discrete m/z values. In profile mode, only finely sampled m/z values above an estimated baseline are stored. In centroid mode, peaks have been located by a simple algorithm in baseline profile spectra. In this work, we assume that the data have been collected in centroid mode. Even though a centroiding algorithm has been applied, a single LC-MS scan will still consist of millions of peaks. In this work, we assume that the data have been collected in centroid mode or converted to centroid mode.

In the first step of processing a single scan, noise peaks are removed by using RemoveNoise(P) below.

RemoveNoise(P). Given a set of peaks, P returns a set of filtered peaks F

\[
T = \text{BuildIndex}(P)
\]

for each p in P

\[
Q = \text{RangeQuery}(T, p.rt - dt, p.rt + dt, \quad \text{p.mz} - \text{dmz}, \text{p.mz} + \text{dmz})
\]

if (size(Q) > R and p.intensity > minI) add p to F

return F

Each peak’s retention time and m/z value are indexed by p.rt and p.mz, respectively. The parameters dt and dmz set the width and height of the planar orthogonal range query. R thresholds the number of peaks, and minI sets the minimum intensity.

Given a set of filtered peaks F, FindXICs() uses planar range queries to construct a graph structure. Connected components in that graph correspond to XICs. Note that the parameters dtF and dmzF set the width and height of the planar orthogonal range query.

FindXICs(F). Given a set of filtered peaks F returns XICs.

\[
D = \text{BuildIndex}(F)
\]

for each p in F

\[
\]

for each returned point q in Q

Add an edge between q and p in a graph G.

return FindConnectedComponents(G)

We made the following modifications in our implementation of the algorithm above:

- We removed XICs that were the result of incorrect centroiding at neighboring m/z values of high-intensity XICs. We use a depth-first approach to finding connected components that does not require explicit construction of the graph.
- We restated precursor intensity of an MS/MS peak by using surrounding peaks in an XIC.
- All intervals used were expressed in parts-per-million (PPM) where m/z \( \times \) ppm \( \times 10^{-6} = Da \)
- We allowed the user to specify m/z ranges of sample contaminants where peaks are filtered out in the LC-MS scan.

To further show the practical utility of using a spatial data structure, we include a function Draw() that illustrates how the spatial data structure T can be used to efficiently draw data points within a region of given width and height on a computer screen. This function plays a key role in allowing an experimenter to assess the quality of the data and set algorithm parameters.

Draw(T, width, height, rt1, rt2, mz1,mz2)

\[
Q = \text{RangeQuery}(T, rt1, rt2, mz1, mz2)
\]

a = width/(rt2 - rt1)

b = height/(mz2 - mz1)

for each q in Q, DrawPeak(a(q.rt - rt1), b(q.mz - mz1)).

Once each scan is processed to find XICs, the computer memory required for the peaks and their corresponding spatial data structures can be freed. All subsequent processing for multiple scans and labeled scans occurs on XICs indexed in a spatial data structure by the retention time and m/z dimensions of their most-intense peak.

Multiple scans are handled by AlignAndGroup() below. The algorithm translates all scans to the reference scan to adjust for differences in when data collection was started in both datasets. It then labels XICs with identifiers that indicate which dataset they belong to and combines all of the XICs into one merged dataset.

AlignAndGroup(S)

for each pair (R, C) in S where R is a reference scan

\[
drt = \text{GetTranslation}(C, R)
\]

Translate each XIC in C by drt in the retention time dimension

for each Si in S

Mark all of the XICs in Si as originating from scan i

Add these labeled XICs to the set Z.

return GroupXICs(Z)

Aligning and grouping requires that GetTranslation() uses reciprocal nearest-neighbor queries to determine corresponding XICs. The median difference in retention time between these XICs is used to align a scan to translation to a reference scan R.

GetTranslation(C, R). Computes the median translational difference between scans.

for each XIC x in C

\[
Q = \text{RangeQuery}(C, x.mz - dmzA, x.mz + dmzA, x.rt - drtA, x.rt + drtA) \text{ find nearest XIC b in Q}
\]

\[
Q = \text{RangeQuery}(C, b.mz - dmzA, b.mz + dmzA, b.rt - drtA, b.rt + drtA) \text{ find nearest r in Q}
\]

if x = r save drt = x.rt - b.rt

return median of drt values

Once scans are aligned to translation GroupXICs() uses planar orthogonal range queries to construct a graph connecting XICs in a merged dataset Z. The range queries use the start and end time of the XIC and a user-specified parameter width in the m/z dimension of an XIC. Connected components in this graph correspond to XICs that have been grouped across scans. Instead of using nonlinear parametric alignment, the range queries automatically account for variance in the position of XICs.

GroupXICs(Z). Groups XICs in Z.

\[
T = \text{BuildIndex}(Z)
\]

for each x in Z

\[
Q = \text{RangeQuery}(T, x.start, x.end, x.mz - dwidth, x.mz + dwidth)
\]

for each returned point q in Q

Add an edge between q and x in a graph G.
return FindConnectedComponents(G)

We made the following modifications in our implementation of the multiscan processing algorithms described above:

- For datasets with replicate runs, we apply grouping hierarchically. We group replicates first, eliminating XICs that do not occur in a sufficient number of replicates.
- We identified ambiguous grouping of XICs when two or more XICs from the same scan occurred in a group. All of the XICs in this group for this scan were removed.

In IsotopePairs(), isotope-labeled data are handled by grouping XICs within a single scan to find light and heavy isotope pairs. The algorithm iterates through XICs by increasing value in the \( m/z \) dimension, pairing light XICs with XICs caused by the heavier species first. By building a spatial index on the XICs, the algorithm assures the same XICs are returned by reciprocal planar orthogonal range queries by using the given label shift in the \( m/z \) dimension.

**IsotopePairs(X).** Finds isotope pairs in a given XIC set X

\[
T = \text{BuildIndex}(X) \\
\text{Sort-by-increasing-m/z}(X) \\
\text{for each } x \text{ in } X \\
\quad \text{if } x \text{ has been paired then skip} \\
\quad Q = \text{RangeQuery}(T, x.\text{start}, x.\text{end}, \\
\quad \quad x.mz + \text{labelshift} - \text{tol}, \\
\quad \quad x.mz + \text{labelshift} + \text{tol}) \\
\quad \text{find largest XIC } x_2 \text{ in } Q \\
\quad Q_2 = \text{RangeQuery}(T, x_2.\text{start}, x_2.\text{end}, \\
\quad \quad x_2.mz - \text{labelshift} - \text{tol}, \\
\quad \quad x_2.mz - \text{labelshift} + \text{tol}) \\
\quad \text{find largest XIC } r \text{ in } Q_2 \\
\quad \text{if } r = x \text{ then save isotope pair } (x, x_2)
\]

We made the following modifications to the algorithm presented above:

- We repeat this process in reverse starting from XICs with large \( m/z \) values to handle the case where the heavy XIC was the only XIC with a fragmentation spectrum.
- We rely on the instrument determined precursor charge to compute isotopic spacing between two XIC pairs.
Fig. S1. Hierarchically clustered protein abundance data. 447 unique proteins were quantified in at least 80 progeny from a cross between \textit{Saccharomyces cerevisiae} strains BY4716 and RM11-1a. Rows correspond to the 447 proteins and columns correspond to the 107 progeny. Nearest-neighbor averaging was used to impute missing data. Yellow designates high and blue designates low abundance relative to the average abundance for the protein.
Fig. S2. (A) Distribution of Pearson correlations between protein abundance values measured between technical and biological replicates in the large Foss et al. dataset. (B) Box plots show the distribution of proteins assayed by Western blot analysis in the original Foss et al. study. GND1, which was present in the original study, was eliminated because we found that the measured tryptic fragment contained a polymorphism.
Fig. S3. Log$_2$-fold change measured by our algorithm and the algorithm used in the Baek et al. study. $r^2 = 0.69$ Spearman's correlation across 1,602 unique protein-coding genes.
Fig. S4.  (A) A planar orthogonal range query determines whether or not a peak is labeled as signal or as noise. A peak is labeled signal if the query returns a threshold number of peaks. (B) After filtering, another set of planar orthogonal range queries are used to connect signal peaks in an undirected graph. XICs correspond to connected components in this undirected graph.
Fig. S5. (A) XICs are labeled according to the scan from which they originate and combined into a single scan. (B) Planar orthogonal range queries are used to connect labeled XIC centers in an undirected graph. The range queries compensate for any nonlinear differences in retention time between LC-MS scans. Connected components in the graph correspond to grouped XICs of the same tryptic peptide.
Table S1. Known femtomole amount of the spiked-in protein injected into each of six datasets collected in triplicate

<table>
<thead>
<tr>
<th>Protein name</th>
<th>Protein injected (fmol) per sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>800</td>
</tr>
<tr>
<td>Carbonic anhydrase</td>
<td>400</td>
</tr>
<tr>
<td>Cytochrome c</td>
<td>200</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>100</td>
</tr>
<tr>
<td>Alcohol dehydrogenase</td>
<td>50</td>
</tr>
<tr>
<td>Aldolase A</td>
<td>25</td>
</tr>
</tbody>
</table>

This table is from the Mueller et al. study.