Journal of proteome research

ARTICLE

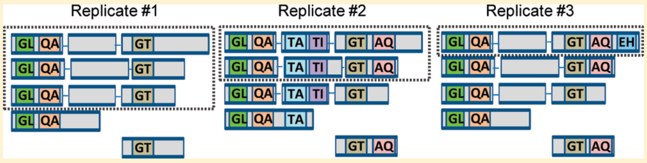
pubs.acs.org/jpr

IsoformResolver: A Peptide-Centric Algorithm for Protein Inference

Karen Meyer-Arendt,[†] William M. Old,[†] Stephane Houel,[†,‡] Kutralanathan Renganathan,[†]
Brian Eichelberger,[§] Katheryn A. Resing,[∥,†] and Natalie G. Ahn[*,†,‡]

[†]Department of Chemistry and Biochemistry and [‡]Howard Hughes Medical Institute, University of Colorado, Boulder, Colorado 80309-0215, United States

Supporting Information



ABSTRACT: When analyzing proteins in complex samples using tandem mass spectrometry of peptides generated by proteolysis, the inference of proteins can be ambiguous, even with well-validated peptides. Unresolved questions include whether to show all possible proteins vs a minimal list, what to do when proteins are inferred ambiguously, and how to quantify peptides that bridge multiple proteins, each with distinguishing evidence. Here we describe IsoformResolver, a peptide-centric protein inference algorithm that clusters proteins in two ways, one based on peptides experimentally identified from MS/MS spectra, and the other based on peptides derived from an *in silico* digest of the protein database. MS/MS-derived protein groups report minimal list proteins in the context of all possible proteins, without redundantly listing peptides. *In silico*-derived protein groups pull together functionally related proteins, providing stable identifiers. The peptide-centric grouping strategy used by IsoformResolver allows proteins to be displayed together when they share peptides in common, providing a comprehensive yet concise way to organize protein profiles. It also summarizes information on spectral counts and is especially useful for comparing results from multiple LC−MS/MS experiments. Finally, we examine the relatedness of proteins within IsoformResolver groups and compare its performance to other protein inference software.

KEYWORDS: proteomics, mass spectrometry, protein inference, peptide-centric, algorithm, spectral counting

INTRODUCTION

An effective method for identifying proteins within complex samples involves multidimensional LC−MS/MS, where proteins are proteolyzed, and peptides are separated by reverse-phase liquid chromatography (RP-LC) and sequenced by mass spectrometry gas phase fragmentation (MS/MS). Automated computer programs are used to analyze the tens of thousands of spectra that can be generated by a single experiment, by matching MS/MS spectra to peptide sequences in protein databases. A significant problem is how to assemble the information contained in large numbers of peptide sequences into a final set of identified proteins.

The task of protein identification is straightforward when peptide sequences are found only within single protein database entries (which we will refer to throughout as "proteins"). However, when a peptide sequence is found in multiple entries, ambiguities arise about which proteins are truly present. This problem is greatest with proteomes where paralogous genes and extensive alternative splicing produce many related proteins within a database.[1] For example, the estimated 20 488 distinct genes in the human genome[2] yield 89 486 proteins in the International Protein Index (v3.75, Aug. 2010) database,[3] which include splice variants, proteolytically processed proteins, and protein fragments. Our analysis shows that of the 3.8 million fully tryptic peptides from this protein database (allowing ≥8 amino acids and up to 2 missed cleavages), over 2 million are shared between two or more proteins. The prevalence of shared peptides creates a need for computational algorithms that infer the most likely protein assignments, a process called protein inference.[4]

Often protein profiles do not report all possible proteins, but only the minimal list which best accounts for the observed peptides (Table 1). The manner in which minimal list proteins are selected differs between protein inference programs. DTA-Select identifies proteins using a greedy algorithm,[5] and in ambiguous cases, shows all possible proteins, allowing users to manually decide between them. ProteinProphet ranks proteins according to probabilities computed from the number of peptides, confidence in the peptide sequence, and the degree to which peptides are shared between multiple proteins.[6] Proteins that are "indistinguishable" (i.e., represented by a set of identical peptides) are assigned equal probabilities. DBParser also uses a greedy algorithm to rank proteins according to those with the most peptides.[7] Phenyx selects a minimal list of proteins, ranked by the number of peptides identified and the protein sequence coverages,[8] but differs from other programs by reporting only one protein entry and accession number (a representative "anchor" protein), even when two or more proteins are indistinguishable. All of these programs use a "protein-centric" approach of matching peptides directly to protein database entries and reporting peptides within the context of proteins (Figure 1a).

In 2004, we proposed an alternative strategy for protein inference, named IsoformResolver, which generates a list of

Received: January 14, 2011
Published: May 21, 2011

Table 1. Terminology

All possible proteins	The complete collection of proteins from which MS/MS observed peptides could be derived.
Minimal list proteins	The smallest number of proteins from which MS/MS observed peptides could be derived.
In silico-derived (ISD) protein groups	The set of all proteins in the protein database clustered by having one or more peptides from an *in silico* digest of that protein database in common.
MS/MS-derived (MSD) protein groups	The set of all possible proteins clustered by having one or more observed peptides in common.
Primary protein	Within an MSD protein group, a protein which has been inferred to be in the minimal list.
Secondary protein	Within an MSD protein group, a protein which may be present, but which has been inferred to not be in the minimal list.
Shared peptide	A peptide which matches two or more protein entries in a protein database.
Bridge peptide	A peptide which matches two or more distinguishable primary proteins.

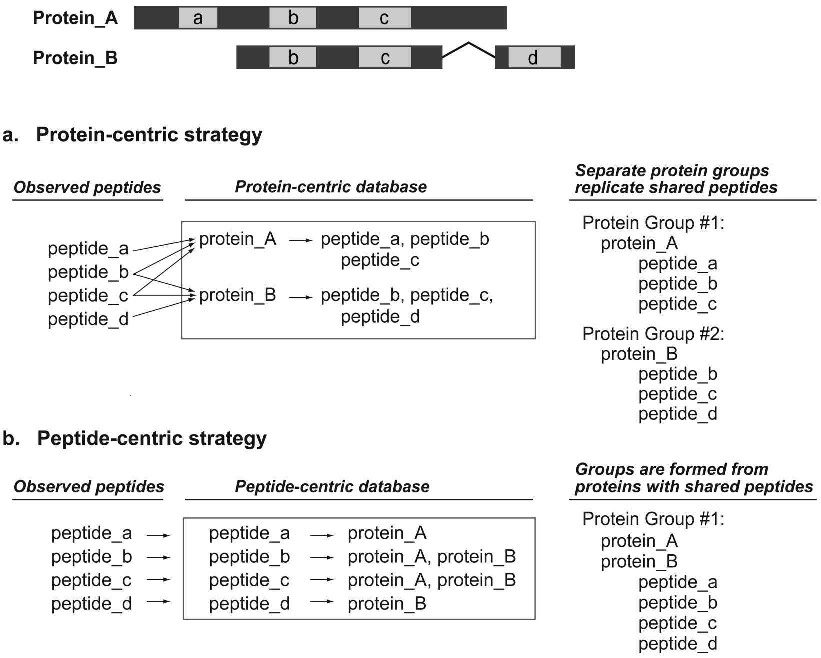


Figure 1. IsoformResolver uses a peptide-centric strategy for protein inference. (a) In a conventional protein-centric approach, observed peptides are searched within a protein sequence database. Protein-centric protein groups replicate peptides when those peptides are found in more than one protein. (b) In the peptide-centric approach, a database consisting of nonredundant peptide sequences is generated from a protein sequence database, where each peptide is matched to all proteins containing the peptide sequence. Observed peptides are matched one-to-one against the list of nonredundant peptide sequences in the database. This allows easy clustering of protein groups that share peptides in common.

nonredundant peptide sequences, and then matches each peptide to all protein entries which contain that sequence.[9] Thus, the approach is "peptide-centric" because the observed peptides are directly referenced against a peptide database (Figure 1b). This strategy has the advantage of more readily assessing the ambiguity in matching peptides to proteins that share peptide sequences in common. Peptides are output within the context of all possible proteins from which they can derive.

In this study, we describe the IsoformResolver algorithm in detail for the first time, and demonstrate the advantages of using peptide-centric protein grouping methods to address problems in protein inference for large data sets. We demonstrate that protein inference increases the variability of proteins between similar data sets ("volatility"), and show that protein inference methods yield significant volatility when reporting proteins separately, which is solved by peptide-centric protein grouping. A compare profile feature of IsoformResolver allows results from many protein profiling experiments to be analyzed, by first performing inference across all experiments pooled together and then reporting spectral counts from individual experiments in an easily viewed format. Finally, we compare IsoformResolver against other protein inference programs and show that the most important factor influencing agreement between different programs is how they treat indistinguishable proteins. Advantages of IsoformResolver are: (i) its protein grouping methods, which allow concise display of proteins including all possible candidates, (ii) its ability to display related proteins adjacently in a protein profile and compare proteomics data sets analyzed at different times and using different software, (iii) its facile integration of label-free quantification by spectral counting into protein sets, and (iv) its ability to compare results from multiple large-scale data sets.

■ METHODS

Data collection and peptide identification

LC−MS/MS data sets used in these studies were collected on human melanoma and erythroleukemia cell lines and summarized

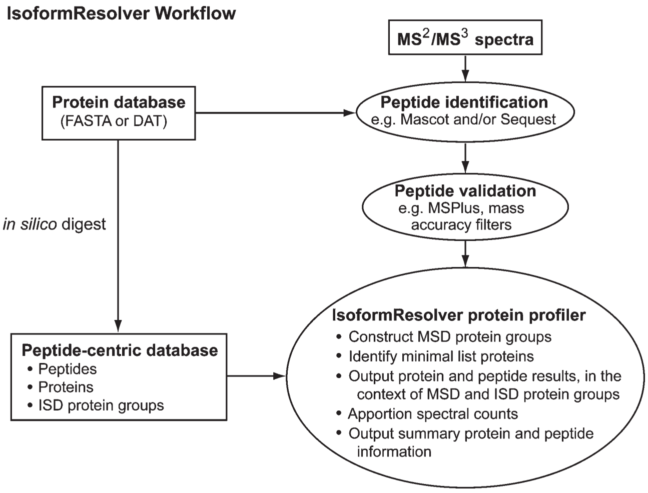

IsoformResolver Workflow
MS^2/MS^3 spectra
Protein database
(FASTA or DAT)
Peptide identification
e.g. Mascot and/or Sequest
Peptide validation
e.g. MSPlus, mass accuracy filters
in silico digest
IsoformResolver protein profiler
• Construct MSD protein groups
• Identify minimal list proteins
• Output protein and peptide results, in the context of MSD and ISD protein groups
• Apportion spectral counts
• Output summary protein and peptide information
Peptide-centric database
• Peptides
• Proteins
• ISD protein groups

Figure 2. IsoformResolver workflow. IsoformResolver inputs a list of experimentally observed peptides identified by a search program, as well as a precalculated peptide-centric database which includes nonredundant peptides, matching proteins, and ISD protein groups. From these input files, IsoformResolver constructs MSD protein groups, identifies primary and secondary proteins, and apportions spectral counts.

in Suppl. Table S1 (Supporting Information). Samples were proteolyzed with trypsin as described,[9−11] and fractionated by reversed-phase HPLC coupled to an LTQ/Orbitrap mass spectrometer (parent scan 475−1600 m/z). DTA files representing MS/MS spectra were generated using BioWorks XCalibur v.3.0 software and concatenated into MGF files using in-house software. DTA files were searched by Sequest[12] specifying carbamidomethylated cysteine and up to two missed trypsin cleavages. Parent ion tolerance was set to 1.2 Da or 50 ppm (specified in Suppl. Table S1) and fragment ion tolerance to 0.8 Da. MGF files were searched using Mascot v.2.2 (Matrix Science,[13]) using the same parameters, and Mascot results were parsed using the Mascot parser (http://www.matrixscience.com/msparser.html). Decoy versions of databases were constructed by reversing each protein sequence from normal databases, which were then searched separately or as a target-decoy database.[14,15] Peptides accepted when scores were above thresholds corresponding to 1% false discovery rate (FDR=FP/(FP+TP)). Peptides were also filtered for physicochemical properties, including peptide size, likely missed cleavages,[16] and mass accuracy (observed minus predicted between −5 ppm and +10 ppm). Peptides were also supported by similarity scoring between observed MS/MS and spectra simulated from peptide fragmentation models[17,18] implemented by Manual Analysis Emulator (MAE).[19]

IsoformResolver Protein Inference Software

IsoformResolver is a Perl program that uses as input one or more files containing validated peptide spectrum matches and generates a protein profile displaying all identified and inferred proteins (Figure 2). For protein information, IsoformResolver accepts any FASTA or EMBL DAT formatted protein databases. Prior to IsoformResolver execution, these protein databases are reformatted into a peptide-centric database, consisting of map files that associate peptides with proteins from which they can be proteolytically derived. This is done once per protein database and requires specifying a protease, number of allowable missed cleavages, and a minimum peptide length. During IsoformResolver execution, validated peptide spectrum matches are input, using the file format shown in Suppl. Figure S1 (Supporting Information). Peptides not found in the peptide-centric database, such as semiproteolytic and nonenzymatic peptides, are searched for within the protein-centric database, and matched to the proteins from which they derive and to the MSD and ISD protein groups to which the proteins belong. Peptides, even semi- and nonproteolytic, are included in all sections of IsoformResolver output and included in spectral counting. Peptide-centric database files have been constructed and tested for use with many proteases including ArgC, LysC, Trypsin, AspN, and can be constructed for any protease with cleavage specificity. In addition, we have constructed and tested peptide-centric database files with combined ArgC + LysC + trypsin cleavages. ISD reformatted datafiles can be constructed from any protein database. The impact of the peptide-centric database will be higher as the number of shared peptides increases. Thus, while ISD protein groups show some benefit using UniProt Sprot, which has a relatively low number of shared peptides, the impact is higher using Sprot/Trembl/Splice variants, a database with an even greater percentage of shared peptides than IPI.

IsoformResolver utilizes two types of protein groups—*in silico*-derived (ISD) protein groups and MS/MS-derived (MSD) protein groups. ISD groups are constructed using all peptides derived from *in silico* proteolysis of a protein database. Using the peptide to protein mapping from the peptide-centric database, proteins are then clustered together whenever they have a peptide in common. Resultant ISD groups are assigned group identifiers and the mapping of proteins to these identifiers are stored in a text file for rapid access during IsoformResolver execution. MSD protein groups are constructed in an identical way, but using different sets of input peptides, consisting of sequences identified from the MS/MS and validated by thresholds or other means. The list of all possible proteins for the observed peptides is obtained by matching peptides to the precalculated peptide-to-protein mapping from the reformatted protein database. These proteins are clustered whenever they have an observed peptide in common, and the resultant protein groups are then assigned an MSD group identifier. MSD groups thus contain only peptides and proteins which were observed in the MS/MS experiment, while ISD groups contain peptides and proteins from the entire protein database, even when they were not observed.

Protein inference is performed on each MSD protein group separately, considering each peptide equally plausible by default, although IsoformResolver can also accept peptide weights using scores or probabilities. Proteins are designated as primary through an iterative process, in which a greedy algorithm is used to select the protein which accounts for the largest number of peptides within a MSD group (or the highest combined score or probability), the protein which accounts for the largest number of remaining peptides that do not match the first protein, and so on until no peptides remain. All other proteins (which lack distinguishing peptide evidence) are designated as secondary. Indistinguishable proteins are primary proteins which are identified by shared peptides that cannot distinguish between the proteins and are counted as a single protein in the minimal list, although all protein identifiers are reported.

In addition to the mapping files described above, the peptide-centric database consists of an annotation file which contains information on the relatedness of proteins within each ISD group. Functional relatedness are evaluated: (i) by gene annotation, based on genes (from Entrez Gene, HGNC, Ensembl, VEGA, or H-InvDB), gene clusters (UniGene) or gene location (chromosomal start location and sense/antisense direction), (ii) by protein family, based on InterPro, Pfam, PROSITE, GENE3D, SUPERFAMILY, PANTHER, ProDOM, PRINTS, and TIGRFAMs databases, and (iii) by GO and other annotations found in the DAT format (e.g., RZPD, UTRdb, SMART, CCDS, CleanEx). Each ISD group has a unique identifier, and is annotated to indicate the percentage of proteins in the group with the same gene, protein family, GO, or other annotation.

Other Protein Inference Programs

Comparisons of IsoformResolver to five other protein inference programs used the following versions of software. Analyses with ProteinProphet[6] used Transproteomic Pipeline (TPP) v.3.3.0 (9/25/2007), and v4.3 JETSTREAM rev 0, Build 200908071234 (MinGW) (http://tools.proteomecenter.org/TPP.php), and were performed using the Mascot option, with peptide probability cutoff 0.95 and protein probability cutoff 0.50. Analyses with Scaffold v.01_07_00 (described in 20 and generously provided by Proteome Software) used the combined Mascot and Sequest option, with peptide and protein probability cutoffs of 0.95 and 0.50, respectively. Analysis with Panoramics v.1 (05/2007, described in ref 21), used the Windows executable provided by the USDA Agricultural Research Service, performed on Mascot search results using protein probability threshold 0.80. IDPicker v.2.0 (described in refs 22 and 23, http://fenchurch.mc.vanderbilt.edu/lab/software.php) used peptide and protein probability cutoffs equal to 0.99. The same Sequest and Mascot results files were used in all analyses, except for IDPicker where data sets were searched using a combined target/decoy database. Analyses with Phenyx Public Server and PhenyxOnline v.2.5 (described in ref 24 and generously made accessible by GeneBio) used the default threshold cutoff (Z-score = 5, p = 0.0001, and AC score = 6).

To compare output between programs, peptides from each program were converted into a common input format, a compare protein profile was created from all peptides generated by the six programs, and the output was annotated with proteins identified by each program. Using IsoformResolver MSD and ISD protein groups, related proteins from each of the profiles were clustered together, simplifying the evaluation in cases where proteins were missed by a profiler or protein variants were identified and allowing for an easy enumeration of primary and secondary proteins.

■ RESULTS

IsoformResolver: Protein Groups and Report Structure

IsoformResolver precalculates a mapping of all proteins to a list of nonredundant peptides within a given database (Figure 2), which identifies all proteins that share peptide sequences. It then generates a protein profile displaying all identified and inferred proteins from one or more files of observed peptides. The peptide-centric algorithm allows two types of protein groups to be generated. *In silico*-derived (ISD) protein groups are constructed from a protein database, by compiling all peptides derived from *in silico* proteolysis (Figure 3a). MS/MS-derived (MSD) protein groups are constructed in an identical way but using input peptides identified experimentally from MS/MS data sets (Figure 3b). Proteins are then assigned to the same group whenever they have a peptide in common. For example, in Figure 3a, proteins_A, _B, _C and _D share peptides and are therefore within the same ISD group. However, proteins_A and _B and proteins_C and _D belong to two MSD groups because not all peptides shared between these proteins are observed. Because only some of all possible peptides can be detected by MS/MS, MSD protein groups are strict subsets of ISD protein groups.

IsoformResolver creates a comma separated values output file which consists of three sections (Figure 4, detailed output in Suppl. Figure S2, Supporting Information). Section 1 displays proteins and peptides within MSD groups, which are in turn listed together within ISD groups. The output catalogues two types of proteins: those that pass Occam's razor test of being among the smallest number that account for the peptide evidence ("primary" proteins), and those that do not ("secondary" proteins). Thus, proteins that account for the greatest number of peptides within an MSD group, or else have distinguishing peptide evidence, are primary; all others are secondary. This nomenclature simplifies, but is nevertheless compatible with, the six protein inference categories previously described.[4,7] Thus, primary proteins include those that are distinct, differentiable, indistinguishable, and proteins identified by shared peptides only when inferred in the minimal list. Secondary proteins include subset, subsumable, and proteins identified by shared peptides only when not inferred in the minimal list.[4] Primary protein identifiers are integral numbers (e.g., 1,2,...) while secondary proteins have alphabetical identifiers (e.g., a,b,...), and common identifiers indicate connectivities between peptides and proteins. For example, in Figure 4, peptides_a, _b, and _c, which match protein_A(identifier 1), will contain "1" in their identifiers. Peptides_b, _c, which match both protein_A(identifier 1) and protein_B(identifier 2), contain both "1" and "2" in their identifiers. Primary proteins that are indistinguishable are marked with an asterisk, for example, peptide_x matches protein_C and protein_D, each with the identifier "3*".

IsoformResolver lists MSD groups in descending order of peptide counts, reporting the observed mass and mass error for each MS/MS, and the number of observed charge forms and highest scores for each peptide, in accordance with reporting guidelines.[25,26] Results from multiple experiments, each containing one or more LC−MS runs, are displayed in separate columns and easily compared using a "compare profile" feature (see below). Section 2 consists of a paragraph summarizing the number of spectra, peptides, proteins, and protein groups, as well as the number of proteins supported by different numbers of peptides. Section 3 summarizes proteins inferred to be in the minimal list in the same order as Section 1 and is in a format that is useful for further automation in spectral count analyses.[10,27,28]

MSD Groups Provide a Complete and Nonredundant Protein Display

Protein inference can be complicated when peptides are shared between multiple protein entries. For example, proteins which are indistinguishable based on the peptide evidence (e.g., proteins_C and _D in Figure 4) complicate the protein report, because the number of proteins in the minimal list (where only one is counted) differs from the number of primary proteins (where both are counted). Reporting all indistinguishable proteins (protein_C and protein_D) inflates the protein count over the minimal list. Selecting one representative protein (protein_C or protein_D) reports the minimum count accurately, but chooses proteins arbitrarily. Treating a set of indistinguishable proteins as one entity with a concatenated name (e.g., protein_C_D) reports

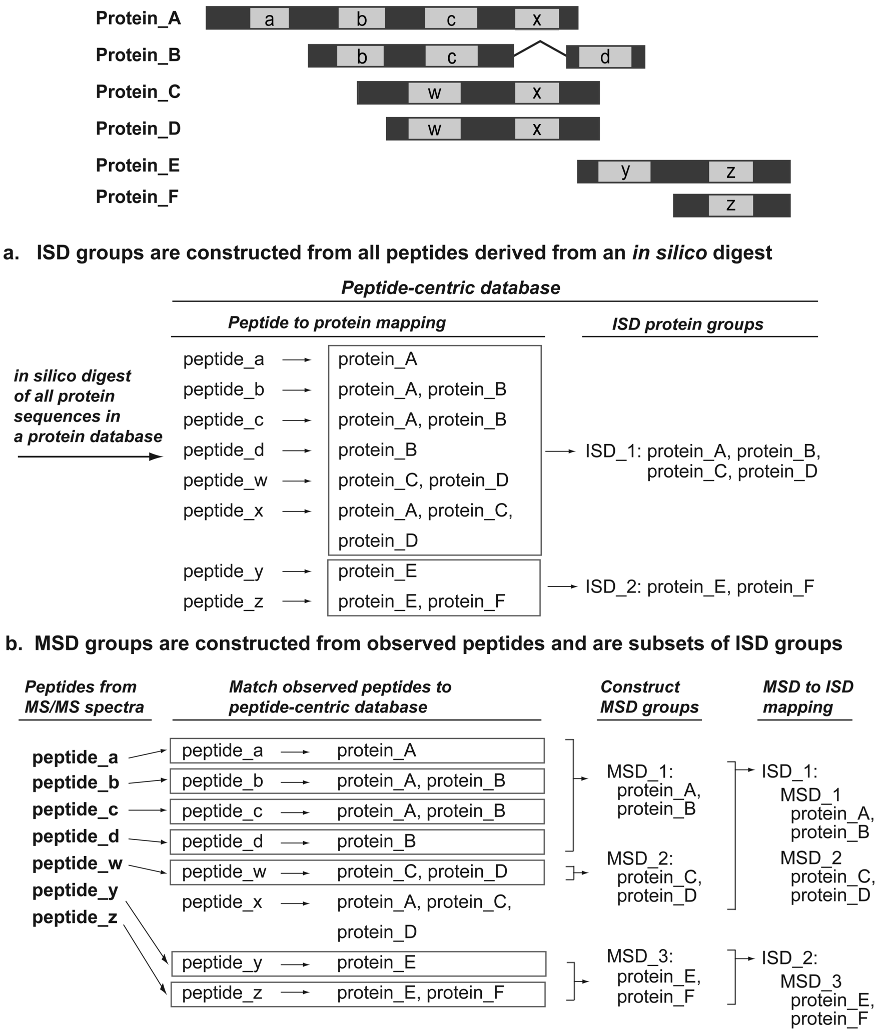

Protein_A
a
b
c
x
Protein_B
b
c
d
Protein_C
w
x
Protein_D
w
x
Protein_E
y
z
Protein_F
z
a. ISD groups are constructed from all peptides derived from an in silico digest
Peptide-centric database
Peptide to protein mapping
ISD protein groups
in silico digest of all protein sequences in a protein database
peptide_a → protein_A
peptide_b → protein_A, protein_B
peptide_c → protein_A, protein_B
peptide_d → protein_B
peptide_w → protein_C, protein_D
peptide_x → protein_A, protein_C, protein_D
→ ISD_1: protein_A, protein_B, protein_C, protein_D
peptide_y → protein_E
peptide_z → protein_E, protein_F
→ ISD_2: protein_E, protein_F
b. MSD groups are constructed from observed peptides and are subsets of ISD groups
Peptides from MS/MS spectra
Match observed peptides to peptide-centric database
Construct MSD groups
MSD to ISD mapping
peptide_a
peptide_b
peptide_c
peptide_d
peptide_w
peptide_y
peptide_z
peptide_a → protein_A
peptide_b → protein_A, protein_B
peptide_c → protein_A, protein_B
peptide_d → protein_B
peptide_w → protein_C, protein_D
peptide_x → protein_A, protein_C, protein_D
peptide_y → protein_E
peptide_z → protein_E, protein_F
MSD_1: protein_A, protein_B
MSD_2: protein_C, protein_D
MSD_3: protein_E, protein_F
ISD_1: MSD_1 protein_A, protein_B MSD_2 protein_C, protein_D
ISD_2: MSD_3 protein_E, protein_F

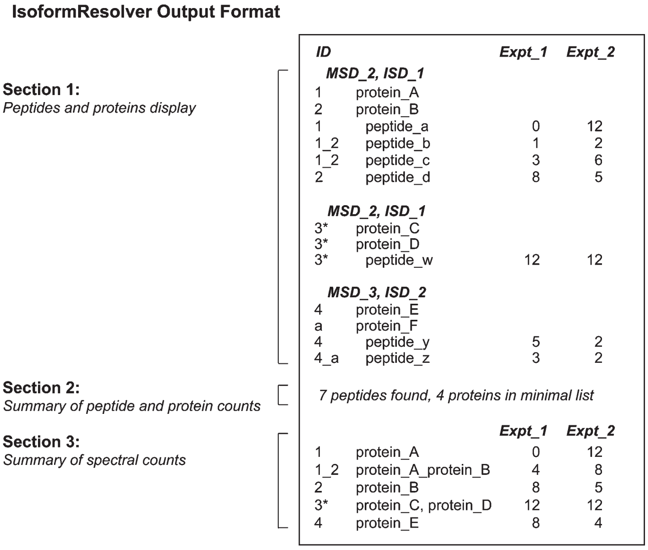
Figure 3. IsoformResolver constructs two kinds of protein groups based on *in silico* derived and on observed peptides. (a) Peptide-centric database enables construction of *in silico* derived (ISD) protein groups, where each ISD group includes proteins that share peptides in common. ISD groups provide a more stable identifier for proteomics results. (b) Experimentally observed peptides are matched to peptide sequences in the peptide-centric database, and proteins are clustered into MS/MS derived (MSD) protein groups when they share observed peptides. MSD groups are subsets of ISD groups and are listed in the output together. Note that peptide_x was not observed experimentally, creating two separate MSD groups within a common ISD group.

the correct number and retains information about the protein identities, but leads to variations in naming between data sets. Each method reports different protein lists, and each compromises accuracy, especially when comparing results from two or more protein profiles.

Also important are cases where peptides are shared between proteins that are distinguishable by the presence of other peptides. We call these cases "bridge peptides" (Our use of the term "bridge peptides" is similar but not identical to the term "razor peptides" (ref 29). The latter refer to peptides which are, by Occam's principle, assigned to the nonoverlapping protein group with the greatest number of peptides. By contrast, bridge peptides are assigned to protein groups which allow overlapping proteins, to retain information that the peptide is shared.), which are shared between primary proteins, and are more problematic than peptides which are shared between primary and secondary proteins. This is because when bridge peptides are encountered by protein-centric inference programs, they are either eliminated from all but one group, or else duplicated and assigned redundantly to different protein groups. An example is shown in the report of two primary proteins, where bridge peptides are replicated and comprise 70% of the peptides for each protein (Suppl. Figure S3a, Supporting Information). Because each protein is listed separately in the output, the replicated peptides may lead to overconfidence in the protein identifications.

Bridge peptides and indistinguishable proteins are a significant problem in protein profiling. For example, in Data set 1A (Suppl. Table S1, Supporting Information), 15% of the 3667 minimal list proteins were linked to others through bridge peptides, 40% were indistinguishable, and only 25% were distinct. Of the 26 225 nonredundant peptides, 67% matched two or more proteins, 7% were bridge peptides, and only 33% matched a single protein entry. Thus, underlying the ambiguity in protein identifications is the fact that the shared and bridge peptides are a considerable fraction of total peptides and affect a high percentage of proteins.

These problems are addressed by IsoformResolver's report format, which lists proteins with shared peptides together, within the context of MSD protein groups. Because primary proteins are


IsoformResolver Output Format
Section 1:
Peptides and proteins display
Section 2:
Summary of peptide and protein counts
Section 3:
Summary of spectral counts

```
ID                          Expt_1   Expt_2
  MSD_2, ISD_1
1      protein_A
2      protein_B
1        peptide_a             0       12
1_2      peptide_b             1        2
1_2      peptide_c             3        6
2        peptide_d             8        5

  MSD_2, ISD_1
3*     protein_C
3*     protein_D
3*       peptide_w            12       12

  MSD_3, ISD_2
4      protein_E
a      protein_F
4        peptide_y             5        2
4_a      peptide_z             3        2

7 peptides found, 4 proteins in minimal list

                            Expt_1   Expt_2
1      protein_A               0       12
1_2    protein_A_protein_B     4        8
2      protein_B               8        5
3*     protein_C, protein_D   12       12
4      protein_E               8        4
```

Figure 4. IsoformResolver output. IsoformResolver output is a comma-separated values spreadsheet file consisting of three main sections. Section 1 lists all possible proteins present in a data set, organized by MSD and ISD groups. Proteins inferred as primary are assigned integral numeric identifiers (e.g., 1, 2...), while secondary proteins are assigned alphabetic identifiers (e.g., a, b...). Peptides are mapped to proteins, using concatenated identifiers (e.g., 4_a) when peptides are shared between more than one protein. Bridge peptides are readily identified as those which map to two or more primary proteins (e.g., 1_2). Indistinguishable proteins are identified by asterisks (e.g., 3*). MSD groups are listed adjacently when they occur within the same ISD group, and are otherwise sorted in descending order by numbers of peptides. Multiple lists of observed peptides can be displayed in a compare profile mode (e.g., Expt_1, Expt_2), allowing easy comparison of proteins and spectral counts between different LC-MS/MS data sets. Section 2 summarizes information on the number of peptides and proteins in the minimal list. Section 3 displays concise information for all proteins in the minimal list and, on bridge peptide regions (on separate lines, marked with concatenated identifiers, e.g., 1_2). Spectral counts for proteins and bridge peptide regions are listed for each experiment. A detailed description of the output can be found in Suppl. Figure S2 (Supporting Information) and an entire output file can be found in Suppl. Worksheet:1.xlsx (Supporting Information).

displayed adjacently when they share peptides, the need to duplicate bridge peptides and redundantly assign them to different proteins is eliminated (e.g., Suppl. Figure S3b, Supporting Information). By displaying all possible proteins, MSD groups allow a user to immediately view the support for inferred proteins as well as alternative but equally likely candidates (Suppl. Figure S2, Suppl. Worksheet:1.xlsx, Supporting Information). The nomenclature used for the MSD identifiers allows the different classifications of distinguishable, indistinguishable, subset, and subsumed proteins to be readily assessed.

ISD Protein Groups Mitigate Volatility Caused by Protein Inference

Problems also arise when protein identifications are easily altered by minor changes in observed peptides, which we refer to as "volatility". Volatility reflects a nonrobust quality of protein inference. Suppl. Figure S4 (Supporting Information) shows an example of assigning peptides to proteins using a greedy algorithm, where two proteins are inferred as primary (IPI00181997.7 and IPI00479677.3), and five proteins are secondary (IPI00376351.2, IPI00383202.1, IPI00744506.1, IPI00785128.2, IPI00797783.1). However, in two equally plausible alternative solutions, IPI00181997.7 and IPI00376351.2 or IPI00376351.2 and IPI00479677.3 could be assigned as the primary proteins. Here, small changes in observed peptides will affect which proteins are deemed primary. For example, if peptide GSL... had not been observed, then IPI00181997.7 would have been inferred as the only primary protein accounting for all peptides, and IPI00479677.3 and IPI00376351.2 would have been called secondary. No method of protein inference obviates this problem, including those which are probability-based, or those which ignore proteins supported by a single peptide.

Protein Repeatability between Replicate Data Sets. To quantify the effects of protein inference on volatility, we examined the repeatability of proteins identified in different data sets, collected at similar depth or varying depth of sampling. First, we quantified the degree to which proteins were repeated between three technical replicate data sets (Suppl. Table S1, Data set 2, Supporting Information), where peptides identified in any data set varied due to random sampling by LC−MS/MS. On average each data set yielded 2922 ± 83 nonredundant peptides (Table 2), 71% of which were found in at least two data sets and 48% which were identical across all three data sets. We then examined all, primary, concatenated, and representative proteins, evaluating their overlap between replicates. As expected, the overlap between replicates was generally higher for proteins than peptides, because each protein was represented by 2.8 peptides, on average. However, we found that the degree of overlap varied with each reporting method (Table 2), due to their differences in how they dealt with indistinguishable proteins.

The overlap was highest when all possible proteins were compared (82% between two or more replicates, 64% between three replicates, Table 2a), because none were removed by inference. In contrast, primary proteins, which listed indistinguishable proteins as separate entities and removed secondary proteins, showed decreased overlap between two replicates (74%) or three replicates (55%), and tended to select for splice variants and proteins that shared many peptides. Concatenated protein identifiers reduced overlap even further (70% between two replicates; 48% between three replicates). Here, indistinguishable proteins were named by concatenated identifiers, which often overlooked proteins present in common between data sets (e.g., an identifier ProteinA_ProteinB would fail to match ProteinB_ProteinC in a different data set, although ProteinB was common to both). Representative proteins increased their overlap between replicates, because proteins with the lowest accession number were chosen from among indistinguishable proteins, while information about other possible proteins was discarded.

Thus, methods which enumerated the most likely proteins (primary and concatenated) paradoxically led to the lowest protein repeatability. Similar trends were observed with proteins identified by two or more peptides (Table 2), indicating that the effect was not caused by peptide sampling variations or low confidence protein identifications. We hypothesized that the effects were instead due to problems introduced by protein inference.

To test this, we constructed a protein profile using a data set which pooled the three replicate data sets together (using a two peptide minimum), then annotated the results by those proteins inferred when each data set was analyzed separately. The minimal list for the pooled data set contained 760 proteins, of which 75

Table 2. Protein Inference Reduces Protein Repeatability between Replicates

		protein reporting method				
	nonredundant peptides	all possible proteins	primary proteins	concatenated proteins	representative proteins	ISD protein groups
Proteins identified by ≥1 peptide[a]						
Replicate 1	2931	3204	2207	1015	1015	896
Replicate 2	2997	3331	2284	1059	1059	933
Replicate 3	2839	3231	2229	1026	1026	902
Total	3989	3972	2936	1418	1298	1109
Present in 2 or more replicates	71%	82%	74%	70%	78%	81%
Present in all 3 replicates	48%	64%	55%	48%	60%	65%
Proteins identified by ≥2 peptides[b]						
Replicate 1	2512	2177	1184	596	596	516
Replicate 2	2533	2244	1197	595	595	513
Replicate 3	2384	2127	1120	571	571	494
Total	3390	2675	1519	796	741	626
Present in 2 or more replicates	71%	82%	75%	71%	78%	81%
Present in all 3 replicates	48%	63%	55%	50%	59%	62%
Proteins identified by ≥2 peptides from pooled replicate data sets[c]						
Total	3479	2791	1507	760	760	656

[a] Proteins identified by one or more peptides showed low overlap between three replicate data sets, due to the effects of protein inference. Comparing proteins at the level of ISD protein groups counteracts this effect, and more accurately captures differences between the replicates. [b] Requiring a minimum of two charge invariant peptides per protein does not mitigate the protein variation. [c] Pooling the replicate data sets together results in fewer proteins in the minimal list. Data for this panel can be found in Suppl. Worksheet:2.xlsx, Supporting Information.

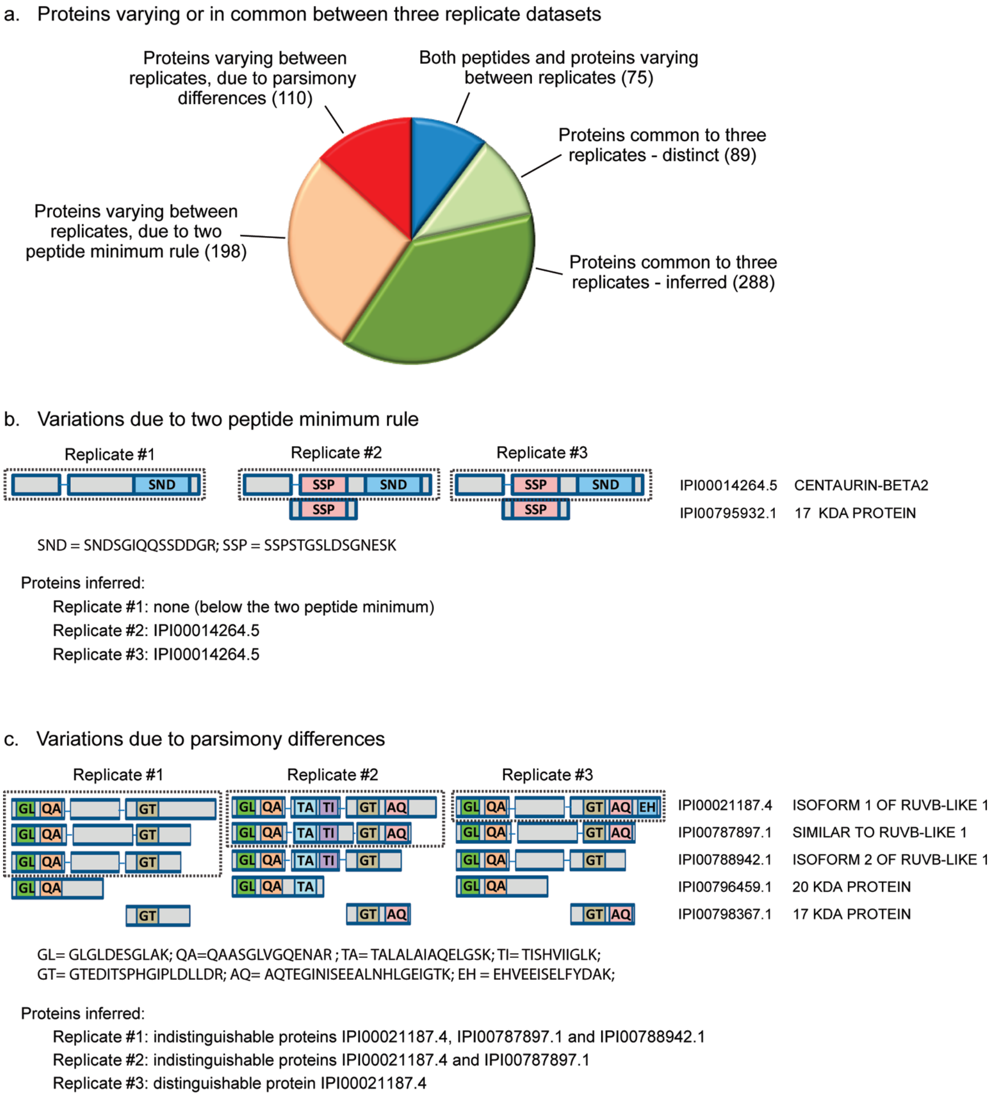
proteins were supported by peptides present in only one or two of the replicates (Table 2, Figure 5a). Thus 685 (90%) of all of proteins were found in common between replicates, far higher than the degree of overlap observed when proteins were inferred from the three data sets independently, regardless of reporting method. Nevertheless, only 377 (55%) of the 685 proteins were inferred in all three replicates (89 distinct proteins, 288 in the same MSD groups), while 308 (45%) proteins differed between replicates. Therefore, the low repeatability across replicate sets was mainly due to variability in the proteins inferred from peptides present in all three sets. In 198 of the 308 cases, the same proteins would have been identified in each data set, but were removed because they were identified by fewer than two peptides (e.g., illustrated in Figure 5b). In the remaining 110 (16%) cases, differences between data sets were due to different protein identifications, and thus caused by parsimony.

Cases where proteins varied due to parsimony reflected volatility due to small changes in additional distinguishable peptides that were present in some, but not all replicates. For example, in Figure 5c, the presence of peptide EH... in Replicate 3 but not Replicates 1 and 2, led to inference of only one primary protein in Replicate 3, whereas three indistinguishable proteins were inferred in Replicate 1 and two indistinguishable proteins were inferred in Replicate 2. Overall, the inferred proteins showed greater differences between replicates than the peptides. These results showed that protein variations are an intrinsic feature of shotgun proteomics, not only due to variations in peptide sampling, but also because variable protein identifications are exacerbated by inference.

We next examined the replicate data sets using ISD protein groups. When each of the three replicate data sets were analyzed separately, 1109 or 626 ISD groups were respectively identified after requiring ≥1 or ≥2 peptides/protein (Tables 2). The overlap in ISD groups between 2 and 3 replicate data sets were 81 and 62−65%, respectively, comparable to the overlap between all possible proteins, and significantly higher than the overlap between inferred proteins, regardless of reporting method. Thus, ISD groups allow greater overlap to compare proteins between data sets, and therefore offers a more stable view of the protein profile.

Protein Repeatability between Data Sets Collected at Different Sampling Depth. Next we examined effects of protein inference on volatility by comparing data sets collected at different depths of sampling, comparing data sets of cell lysate proteins analyzed in duplicate 1D-LC-MS/MS runs (29,907 MS/MS) vs proteins separated by SDS-PAGE followed by in-gel digestion (252,205 MS/MS) (Suppl. Table S1, Data set 3, Supporting Information). Prior studies had shown that proteins identified in data sets at lower sampling depth overlap nearly completely with those in data sets collected at higher depth.[9] Thus as expected, the overlap was high, where 91% of peptides and 98% of proteins identified in the lower sampling depth data set were also identified in the higher depth data set (Table 3a). However, the overlap between primary proteins was only 75%.

To confirm that this variability was due to inference and not to differences in peptides between the peptides contained in each data set, we simulated a lower depth data set by truncating MS/MS spectra with lowest intensity from the higher depth Data set 3. The MS/MS removed were adjusted to yield a remaining number of peptides similar to that of the lower depth experimental data set (Table 3). Because peptides in the truncated data set were a complete subset of those in the high depth data set, any protein variations would reveal effects due only to inference. The results showed that even when the peptides in the low depth data set overlapped those in the high depth data set completely,


a. Proteins varying or in common between three replicate datasets

Proteins varying between replicates, due to parsimony differences (110)
Both peptides and proteins varying between replicates (75)
Proteins common to three replicates - distinct (89)
Proteins varying between replicates, due to two peptide minimum rule (198)
Proteins common to three replicates - inferred (288)

b. Variations due to two peptide minimum rule

Replicate #1
Replicate #2
Replicate #3
SND
SSP
IPI00014264.5 CENTAURIN-BETA2
IPI00795932.1 17 KDA PROTEIN

SND = SNDSGIQQSSDDGR; SSP = SSPSTGSLDSGNESK

Proteins inferred:

Replicate #1: none (below the two peptide minimum)
Replicate #2: IPI00014264.5
Replicate #3: IPI00014264.5

c. Variations due to parsimony differences

Replicate #1
Replicate #2
Replicate #3
GL
QA
TA
TI
GT
AQ
EH
IPI00021187.4 ISOFORM 1 OF RUVB-LIKE 1
IPI00787897.1 SIMILAR TO RUVB-LIKE 1
IPI00788942.1 ISOFORM 2 OF RUVB-LIKE 1
IPI00796459.1 20 KDA PROTEIN
IPI00798367.1 17 KDA PROTEIN

GL= GLGLDESGLAK; QA=QAASGLVGQENAR ; TA= TALALAIAQELGSK; TI= TISHVIIGLK;
GT= GTEDITSPHGIPLDLLDR; AQ= AQTEGINISEEALNHLGEIGTK; EH = EHVEEISELFYDAK;

Proteins inferred:

Replicate #1: indistinguishable proteins IPI00021187.4, IPI00787897.1 and IPI00788942.1
Replicate #2: indistinguishable proteins IPI00021187.4 and IPI00787897.1
Replicate #3: distinguishable protein IPI00021187.4

Figure 5. Protein repeatability is affected by protein reporting methods and volatility. (a) Three pooled replicate data sets analyzed separately showed varying levels of agreement between identified proteins. Of 760 proteins identified when replicate data sets were analyzed together, 685 (all but 75) were inferred from peptides present in all 3 replicates, signifying high peptide overlap between data sets. Of these, only 377 (55%) proteins were present in all three replicates, revealing low protein overlap. (b) Examples of protein variations between replicates introduced by protein inference. Replicate 1 in Data set 2 shows only one peptide and is therefore not matched to a protein, while Replicates 2 and 3 infer IPI00014264.5 from two peptides. (c) Three of seven peptides are found in all three replicates, but peptides in Replicate 1 cannot distinguish between IPI00021187.4, IPI00787897.1 and IPI00788942.1, and peptides in Replicate 2 cannot distinguish IPI00021187.4 and IPI00787897.1. In contrast, peptide EH... identifies IPI00021187.4 as the sole primary protein in Replicate 3. The full output can be viewed in Suppl. Worksheet:2.xlsx (Supporting Information).

protein inference decreased the overlap between primary proteins by 21%.

By contrast, ISD groups showed 98% overlap between data sets collected at lower and higher sampling depth and retained 100% overlap between the simulated and higher depth data sets. Thus, the mapping of proteins and peptides to invariant ISD groups added stability to the protein report, bypassing problems in reproducibility, and thereby counteracting volatility caused by protein inference.

Compare Profile Feature Optimizes Information Retrieval from Multiple Experiments

We found that protein inference varied when data sets were joined in different ways. Often, proteomics experiments involve comparisons between LC−MS/MS runs (e.g., control vs treated, differing protocols, chromatographically separated proteins). The many data sets produced can be analyzed either carrying out protein inference on each data set separately and then combining the results to create an aggregate set ("aggregate" analysis), or by pooling peptides from all data sets together before protein inference ("pooled" analysis) (Figure 6a).

In order to compare the two approaches, data sets were collected on cell lysate proteins that were first separated into 33 fractions by strong anion exchange (SAX) chromatography, followed by proteolysis and LC−MS/MS (Suppl. Table S1, Data set 1B, Supporting Information). In a first test, proteins were assembled from data sets of each fraction analyzed separately by IsoformResolver, which were then joined into an aggregate profile of 7699 primary (distinct + distinguishable + total indistinguishable) proteins and 4582 minimal list

Table 3. Protein Inference Methods Underestimate Overlap between Proteins Obtained from Data Sets with Lower vs Higher Depth of Sampling

			protein reporting method			
	MS/MS spectra	nonredundant peptides	all proteins	primary proteins	representative proteins	ISD protein groups
Comparison of lower depth (unfractionated) vs higher depth (fractionated) data sets[a]						
Lower depth data set	29 907	660	3631	2260	941	836
Higher depth data set	252 205	11 112	8502	5064	2402	2098
Overlap (as % of lower depth data set)	N.A.	91%	98%	75%	91%	98%
Comparison of simulated lower depth vs higher depth data sets[b]						
Simulated lower depth data set	24 338	3663	3895	2448	1052	935
Higher depth data set	252 205	11 112	8502	5064	2402	2098
Overlap (as % of lower depth data set)	N.A.	100%	100%	79%	94%	100%

[a] Unfractionated lysates are sampled at a lower depth, while fractionated lysates allowed for higher sampling depth. [b] Lower depth of sampling is simulated by removing >90% of the lowest intensity MS/MS from the higher depth data set. Even when the peptide overlap between lower and higher depth data sets is 100%, different proteins were inferred.

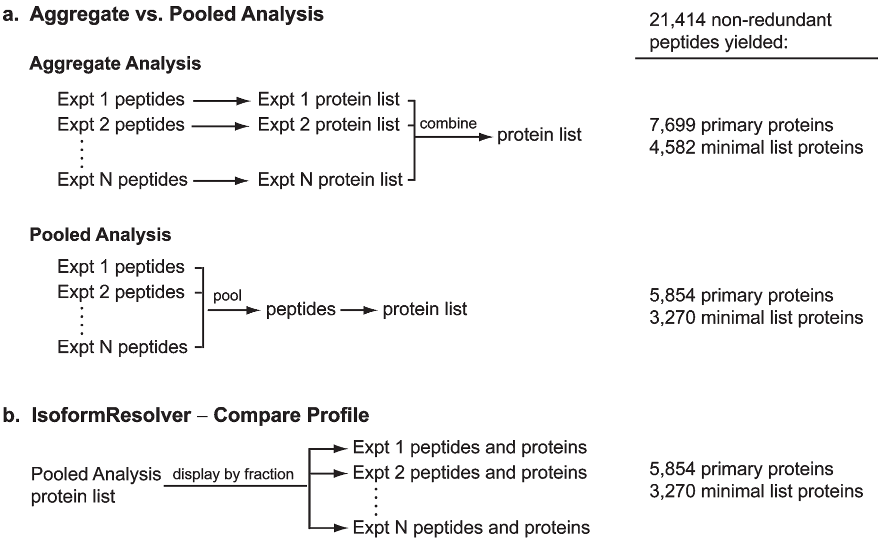


Figure 6. Protein inference is affected by joining multiple data sets in different ways. (a) In an aggregate analysis, peptides from LC−MS/MS data sets of different fractions from a chromatographically resolved sample are first analyzed by inference, then the proteins are combined. In a pooled analysis, peptides from different fractions are combined prior to protein inference. Pooling peptides and then performing inference yields the simplest solution with the smallest number of proteins, as shown at the right (Data set 1B), but loses important information when analyzing fractions separately. (b) In a compare profile, IsoformResolver combines the strengths of pooled and aggregate analyses, by pooling the data sets to identify the minimal list proteins, and then displaying spectral counts for each individual data set.

(distinct + distinguishable + minimal indistinguishable) proteins, where the counting excluded redundant cases. In a second test, peptides from each SAX fraction were combined into one pooled data set and then assembled into proteins using IsoformResolver, yielding 5854 primary and 3270 minimal list proteins. Thus, the number of minimal list proteins inferred in the pooled profile was 40% lower than those inferred in the aggregate profile. The protein overlap was nearly complete, as only one primary protein observed in the pooled analysis was excluded from the aggregate analysis. Therefore, with multiple LC−MS/MS runs, pooling the peptide information before assembly yielded a more conservative protein count.

How protein inference underlies this effect is illustrated in an example where 6 observed peptides mapped to 6 possible proteins (Suppl. Figure S5, Supporting Information). In the pooled analysis, two primary proteins (IPI00444788.1 and IPI00025340.3) accounted for all peptides (Suppl. Figure S5a, Supporting Information). However, in the aggregate analysis, the number of peptides in each fraction varied, and together inferred six primary proteins, five of which were distributed in three indistinguishable sets (Suppl. Figure S5b, Supporting Information). For example, peptides in fraction #22 identified four indistinguishable proteins (IPI00444788.1, IPI00445123.1, IPI00456744.1 and IPI00743804.1), while peptides appearing in fraction #23 identified two indistinguishable proteins (IPI00444788.1 and IPI00456744.1). Thus, even when the same peptides were represented, carrying out protein inference on separate data sets inflated the protein counts compared to pooling the data sets prior to inference. Such differences were caused by lower numbers of peptides in each fraction in the

a. MSD protein group (GNPDA1, GNPDA2)

Prot ID	Accession	Protein descriptor	Gene
1009*	IPI00550894.4	GLUCOSAMINE-6-PHOSPHATE ISOMERASE SB52	GNPDA2
1009*	IPI00744859.1	GLUCOSAMINE-6-PHOSPHATE DEAMINASE 2 (FRAGMENT)	GNPDA2
1010	IPI00009305.1	GLUCOSAMINE-6-PHOSPHATE ISOMERASE	GNPDA1

Prot ID	Peptide sequence	Highest XCORR	Highest Mowse	Spectral Counts	
1009*	EAGGIDLFVGGIGPDGHIAFNEPGSSLVSR	4.5	99.1	3	9
1009*	NHPESYHSYMWNNFFK	2.8	30.5	6	
1009*_1010	AIEEGVNHMWTVSAFQQHPR	6.2	96.3	22	45
1009*_1010	EVMILITGAHK	3.2	50.7	2	
1009*_1010	TFNMDEYVGLPR	3.9	82	14	
1009*_1010	VPTMALTVGVGTVMDAR	4.6	117.2	7	
1010	AAGGIELFVGGIGPDGHIAFNEPGSSLVSR	5.6	94.6	10	37
1010	DHPESYHSFMWNNFFK	5.4	99.1	8	
1010	FFDGELTK	2.1	52	2	
1010	LIILEHYSQASEWAAK	5.4	70	13	
1010	LVDPLYSIK	2.7	35.6	4	

b. Apportionment of bridge peptide spectral counts

Protein ID	Accession	Gene	ISD group	Spectral counts (SC)	App SC	Peptides	App Peptides
1009*	IPI00550894.4	GNPDA2	ISD8-4884	9	17.8	2	2.8
1009*_1010	IPI00550894.4_IPI00009305.1	GNPDA2	ISD8-4884	45		4	
1010	IPI00009305.1	GNPDA1	ISD8-4884	37	73.2	5	8.2

c. Spectral count tables for fractionated datasets

Case [i]

								Fractions			
Protein ID	Accession	Gene	ISD group	SC	App SC	Pep	App Pep	19	20	21	22
1009*	IPI00550894.4	GNPDA2	ISD8-4884	9	17.8	2	2.8	2	7	0	0
1009*_1010	IPI00550894.4_IPI00009305.1	GNPDA2	ISD8-4884	45		4		8	23	10	4
1010	IPI00009305.1	GNPDA1	ISD8-4884	37	73.2	5	8.2	3	21	9	4

Case [ii]

								Fractions									
Protein ID	Accession	Gene	ISD group	SC	App SC	Pep	App Pep	14	15	16	17	18	19	20	21	22	23
1065	IPI00002519.1	SHMT1	ISD8-207	103	108.2	13	13.6	5	27	38	28	3	2	0	0	0	0
1065_1066*	IPI000025191_IPI00002520.1	SHMT2	ISD8-207	8		1		2	2	2	2	0	0	0	0	0	7
1066*	IPI00002520.1	SHMT2	ISD8-207	57	59.8	6	6.4	0	0	0	9	1	9	8	13	10	7

Case [iii]

								Fractions											
Protein ID	Accession	Gene	ISD group	SC	App SC	Pep	App Pep	28	29	30	31	32	33	34	35	36	37	38	39
248	IPI00641829.5	BAT1	ISD8-184	74	147.6	11	17.1	0	20	46	8	0	0	0	0	0	0	0	0
248_249	IPI00641829.5_IPI00644431.1	DDX39	ISD8-184	107		9		0	28	43	19	6	5	3	2	0	0	1	0
249	IPI00644431.1	DDX39	ISD8-184	19	52.4	5	7.8	0	6	6	5	2	0	0	0	0	0	0	0

Case [iv]

								Fractions			
Protein ID	Accession	Gene	ISD group	SC	App SC	Pep	App Pep	35	36	37	38
121*	IPI00003964.3	USP9X	ISD8-1819	64	152.1	20	2.8	0	17	37	27
121*_122*	IPI00003964.3_IPI00012094.1	USP9Y	ISD8-1819	72		18		3	12	40	17
122*	IPI00012094.1	USP9Y	ISD8-1819	1	1.9	11	8.2	0	0	1	0

Case [v]

							Fractions					
Protein ID	Accession	Gene	ISD group	Total SC	Pep	App Pep	27	28	29	30	31	32
1470*	IPI00258833.1	SNX6	ISD8-1020	16	4	4	2	4	7	3	0	0
1290	IPI00295209.5	SNX5	ISD8-1020	38	8	8	0	0	5	23	8	2

Figure 7. Apportionment of bridge peptides for spectral counting. (a) An example shows 11 peptides observed in data sets of 33 fractions, mapped to three glucosamine-6-phosphate isomerase/deaminase proteins (GNPDA1, GNPDA2). Four peptides bridge both sets of primary proteins (IPI00009305.1 and the indistinguishable set of IPI00550894.4 and IPI00744859.1). (b) The spectral count summary shows that bridge peptides account for nearly half of the spectral counts. Spectral counts for the bridge regions are apportioned to each primary protein, proportional to the spectral counts for their distinguishing peptides. (c) Using a compare profile to report spectral counts across multiple data sets facilitates evaluation of primary proteins and apportionment of bridge peptides. **Case** [**i**] Bridge peptides track each of two primary proteins across fractions. **Case** [**ii**] Bridge peptides track with and are more accurately apportioned to IPI00002519.1, with minor overlap with IPI00002520.1. **Case** [**iii**] Bridge peptides provide evidence for a protein that likely differs from Bat1 and Ddx39. **Case** [**iv**] Bridge peptides track IPI0003964.3 and IPI00012094.1, but are more accurately apportioned to IPI0003964.3, given low spectral count evidence for IPI00012094.1. **Case** [**v**] No bridge peptides were observed in the data set, although IPI00258833.1 and IPI0295209.5 are related proteins. The fact that these proteins are related would have been missed had they not been listed within the same ISD protein group.

aggregate analysis, leading to increased numbers of indistinguishable proteins. In the pooled analysis, more proteins were converted to distinguishable or secondary proteins, reducing the indistinguishable proteins and minimizing the number of primary proteins.

Despite this advantage, pooling data sets discarded important information about the representation of different proteins across samples. For example, when chromatographically separating proteins, it is often useful to know how different proteins vary in elution, and here, it would be advantageous to analyze each data set separately. Therefore, IsoformResolver provides the option of displaying a "compare profile" in Section 3 (Figure 6b), in which primary proteins are inferred and spectral counts apportioned using the pooled data sets, while spectral counts are displayed per individual data set.

An example of a compare profile is shown in Suppl. Figure S5c (Supporting Information), where the pooled analysis inferred two primary proteins, and displaying each fraction separately in the output clearly showed that the two proteins resolved chromatographically. Peptides in fractions #15−17 best matched protein_764, while peptides in fractions #22−24 best matched protein_763 or secondary protein_b. In fact, in fractions #22−24, support for protein_b over protein_763 was suggested by the absence of peptide LEE... against the presence of peptides LSE..., SLS..., SPP... and KLP.... This illustrates the advantage of combining the peptide evidence with information about chromatographic resolution, allowing the user to evaluate cases that might otherwise have been overlooked. By calculating the most conservative estimate of minimal list proteins and displaying related proteins in logical groupings, IsoformResolver allows spectral count variations between individual data sets to be readily evaluated. Thus, the compare profile feature of IsoformResolver combines the strengths of pooled and aggregate analyses, by providing a conservative calculation of proteins from a pooled analysis and an informative display of results in each experiment.

IsoformResolver Simplifies the Spectral Count Analysis of Bridge Peptides

An important approach for label-free quantification of proteins is spectral counting, which sums the total number of MS/MS corresponding to any peptide in a given protein.[27] However, assigning spectral counts to proteins is complicated when bridge peptides are shared between two or more proteins in the minimal list.[30,31] This can skew information on relative abundances of proteins. For example, in Figure 7a, 2 peptides (EAG..., NHP...) uniquely infer two indistinguishable proteins (GNPDA2) with 9 spectral counts, and 5 peptides (AAG..., DHP..., FFD..., LII..., and LVD...) uniquely infer one protein (GNPDA1) with 37 spectral counts. Four bridge peptides (AIE..., EVM..., TFN..., VPT...) represent an additional 45 spectral counts, and how these are apportioned can greatly influence the estimated relative abundance of GNPDA1 and GNPDA2. IsoformResolver apportions spectral counts from bridge peptides proportionally to the spectral counts of nonshared peptides for distinguishable proteins. In this example, 20% of spectral counts from bridge peptides were apportioned to GNPDA2 and 80% were apportioned to GNPDA1 (Figure 7b). Similar calculations are used to apportion nonredundant peptides. Apportioned spectral counts for bridge and nonredundant peptides are then summarized in Section 3 of the IsoformResolver output (Figure 4, Suppl. Figure S2, Supporting Information). We report spectral counts for distinguishable and bridge peptides separately, as the primary evidence for each protein. Apportionment of spectral counts according the number of distinguishable peptides is also included which can be useful for comparing proteins containing bridge peptides with those that do not.[30,31]

Figure 7c shows examples which break down spectral counts according to SAX fractions, and illustrate how spectral counts for nonbridge vs bridge peptides can provide information about the reliability of protein identifications and the presence of related proteins. **Case** [**i**] shows a simple example, where bridge peptides track two proteins (1009*, 1010) in each of fractions #19−22, and support the presence of each protein. **Case** [**ii**] shows bridge peptides which match two primary proteins (1065, 1066*) but track only one protein (1065). In **Case** [**iii**], some bridge peptides appear in fractions #33−39 but track neither primary protein (363or364), suggesting that they instead correspond to another protein. Because IsoformResolver reports detailed information about all proteins and their spectral counts, such cases can be readily assessed and overlooked proteins identified.

A unique feature of IsoformResolver is that it clusters the display of proteins based on shared peptides, allowing proteins related by bridge peptides and belonging to the same MSD and ISD groups to be listed adjacently. This solves problems caused by listing proteins separately, which may lead to overconfidence in protein identifications. For example, Figure 7c, **Case** [**iv**] shows two paralogous proteins from different genes which differ widely in spectral counts (81 for 121*, 1 for 122*). Redundantly assigning the 18 bridge peptides to both proteins might create false confidence for the presence of protein 122*, especially if the proteins were reported in different regions of the output. By displaying these proteins adjacently in the output, potential false positive peptide assignments (e.g., with disproportionately few spectral counts) and the apportionment of bridge peptides are readily evaluated. In addition, clustering proteins by ISD groups allows related proteins to be easily identified. For example, in **Case** [**v**], proteins 1470* and 1290 are paralogs that share amino acid sequences, but no bridge peptides were observed and the peptides for proteins 1470* and 1290 were nonoverlapping. Here, protein-centric methods would have placed each protein in separate groups, and the fact that these genes are related would have been missed. The ability of IsoformResolver to display ISD groups adjacently allowed these related gene products to be listed together, facilitating evaluation of their relative abundance by spectral counting.

Proteins in ISD Groups Are Functionally Related and Vary with Shared Peptide Length

We evaluated whether ISD groups might contain proteins that share biological function as well as peptide sequence. Functional relatedness was evaluated in multiprotein ISD groups (i.e., with two or more proteins), scoring agreement between IPI UniProt (DAT) and GO database annotations, and requiring one or more annotation to be shared in common among all proteins within an ISD group. We assessed first whether proteins within each group were derived from common genes; second, whether they were members of a common protein family, although not derived from a common gene; and third, whether they were functionally related by GO or other annotations, although not a common protein family.

Of the 10 651 multiprotein ISD groups generated from shared peptides of 8 amino acids or longer, 7136 (67%) contained protein members all derived from a common gene (e.g., splice variants, processed protein forms), 1683 more (16%) contained members all belonging to a common protein family, and 538

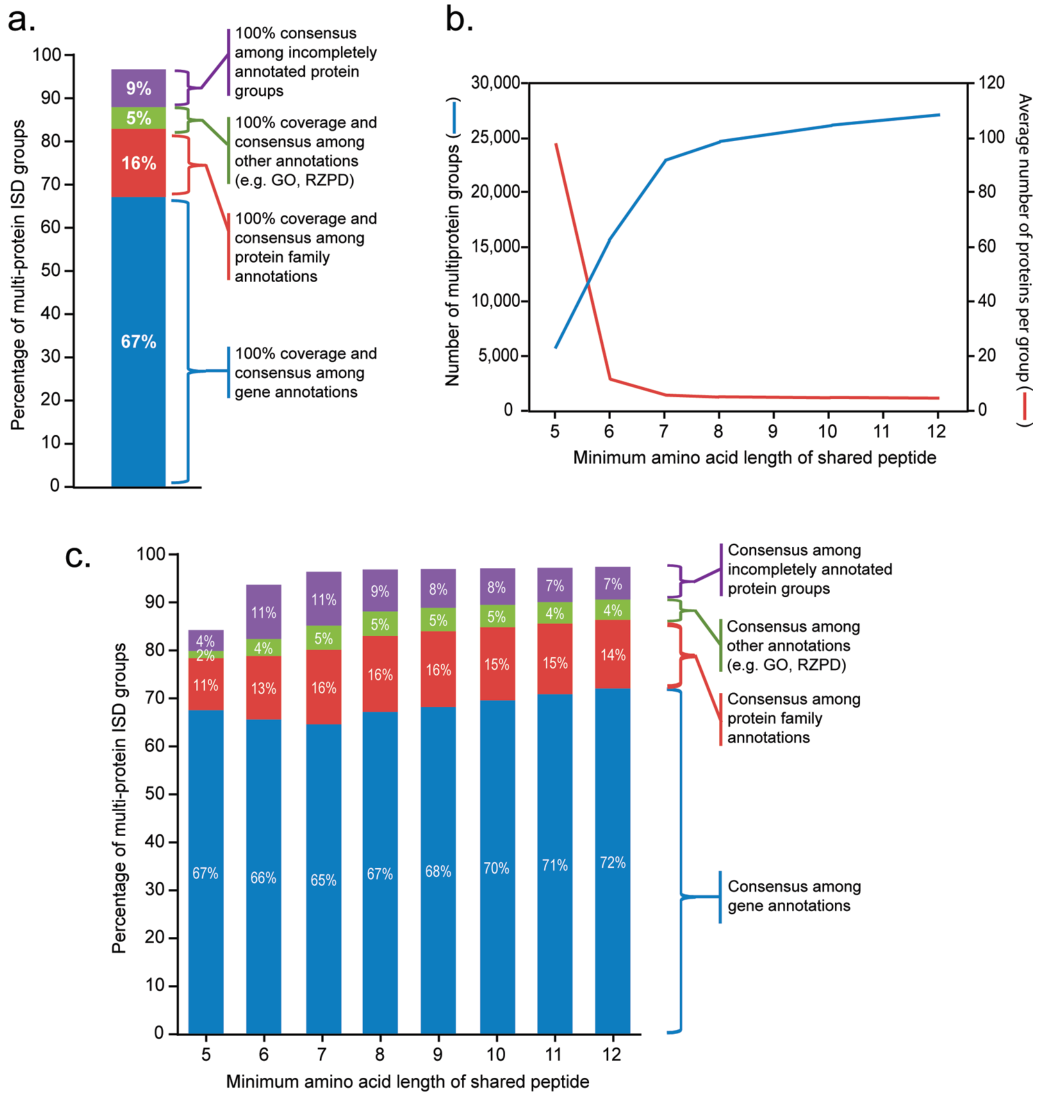


Figure 8. ISD protein groups define functionally related proteins. (a) Proteins show strong functional relatedness within multiprotein ISD groups (defined by a minimum shared peptide length of 8 amino acids). In 67% (7136 of 10 630) ISD groups where all proteins are annotated, all proteins within the group were related based on gene database annotation and agreement on a gene identifier. In 16% (1684) of cases, all proteins within the group were related based on protein family annotation, and consensus for a single identifier. In 5% (538) of cases, all proteins were related based on GO or other database annotations. In ISD groups where the proteins were incompletely annotated, 9% (929) showed complete agreement in gene, protein family, GO or other annotations for those proteins which were annotated. (b) As the minimum length of the shared peptides are set to increasing values, the number of ISD protein groups increases while the average number of proteins per ISD protein group decreases. (c) Consensus and functional relatedness between proteins within each ISD group increases as the length of shared peptides increases.

more (5%) contained protein members sharing GO or another cross-reference annotations (Figure 8a). Another 929 (9%) contained members with incomplete annotations; however, the proteins that were annotated showed complete agreement in gene, protein family, or other annotations. Thus in 97% of ISD groups, all protein members that could be evaluated were functionally related. In the remaining 3%, proteins often appeared related. For example, one group contained proteins with similar gene names (CNNM1, CNNM2, CNNM3, and CNNM4, corresponding to cyclins M1−M4), even though their annotations were nonoverlapping. Because gene names do not always report function, this group was not scored, although its members were clearly related.

We also examined the frequency with which proteins between different ISD groups were unrelated. Here, we scored "exclusivity", when a group was the only one which corresponded to a particular gene, protein family, or other cross-reference identifier. Among the 7136 ISD groups whose protein members unanimously specified a single gene annotation, 7041 (99%) were exclusive. Not surprisingly, protein family annotations did not show the same degree of exclusivity. Among 5868 groups whose proteins unanimously specified a common protein family annotation (4185 also specifying a common gene), only 1410 cases were exclusive. The results show that proteins that share even few peptides in common are related functionally, and that for the most part, ISD groupings capture all proteins which are related, while excluding proteins which are unrelated.

This behavior changed with the length of shared peptides. Protein groups constructed from shared peptides with minimum

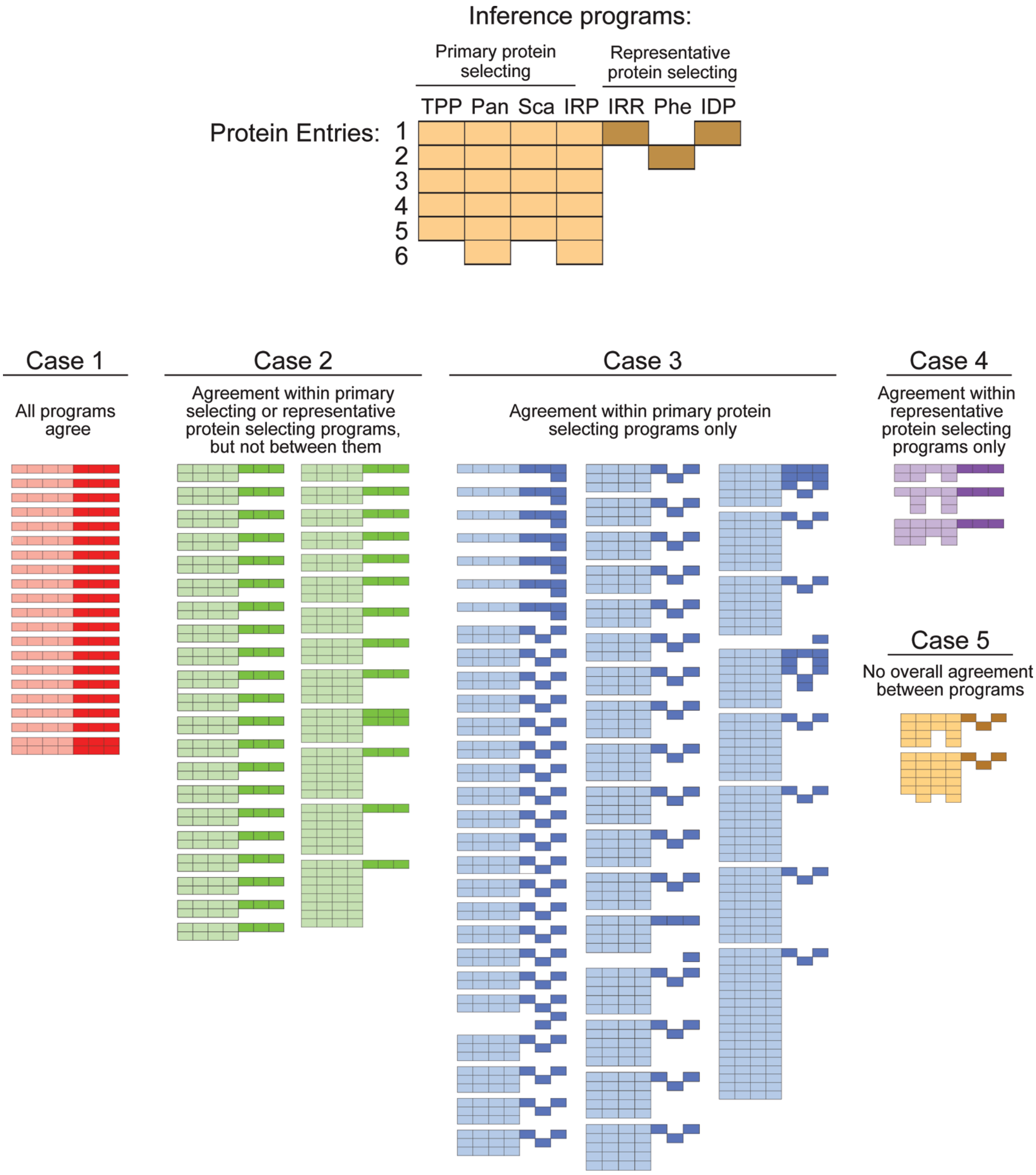

Inference programs:
Primary protein selecting
Representative protein selecting
TPP Pan Sca IRP IRR Phe IDP
Protein Entries: 1 2 3 4 5 6
Case 1
All programs agree
Case 2
Agreement within primary selecting or representative protein selecting programs, but not between them
Case 3
Agreement within primary protein selecting programs only
Case 4
Agreement within representative protein selecting programs only
Case 5
No overall agreement between programs

Figure 9. Protein inference differences are mainly due to whether programs report primary or representative proteins. Six protein inference programs (TPP ProteinProphet, Panoramics, Scaffold, IsoformResolver, Phenyx, and IDPicker) were used to analyze a data set. One-hundred twelve ISD protein groups with common peptides are shown here. Color shadings distinguish programs which report primary proteins (TPP ProteinProphet, Panoramics, Scaffold, IsoformResolver in its default mode), versus programs that report a single representative protein from among indistinguishable proteins (IsoformResolver in its representative selecting mode, Phenyx, and IDPicker). (Top) Close-up of Case 5 illustrating the organization of each case. In this example between 1 and 6 proteins were reported for this ISD protein group, and boxes indicate proteins inferred by each program. (Bottom) all 112 ISD protein groups. In Case 1 (20 of 112 ISD groups), identical proteins were inferred by all programs. In Case 2 (34 groups), proteins agreed between primary selecting programs and between representative selecting programs, although the two types of programs disagreed. In Case 3 (53 groups), primary selecting inference programs agreed with each other, but programs which select representative proteins did not. In Case 4 (3 groups), only representative selecting programs agreed. In Case 5 (2 groups) there was no overall agreement in either set of programs. Additional information is in Suppl. Table S3 and the entire annotated compare profile can be found in Suppl. Worksheet:3.xlsx (Supporting Information).

length 5, 6, or 7 amino acids produced fewer protein groups, each with higher average numbers of proteins (Figure 8b). On the other hand, as peptide length and the number of groups increased, the relatedness of proteins within each group also increased (Figure 8c). Considering only gene annotations, increased peptide length led to increased consensus, while exclusivity remained constant (data not shown). A minimum length of 5 amino acids yielded large ISD groups, averaging 98 protein entries, whose proteins exhibited functional relatedness within 84% of groups. A minimum length of 12 amino acids yielded more ISD groups, with little change in functional relatedness compared to 8 amino acids. Overall, 8 amino acids was the optimal minimum length for grouping proteins with common function. This was the minimum length previously determined for filtering out false positives during peptide identification.[9,14] We conclude that 8 amino acids provide an optimal minimum peptide length for protein grouping as well as peptide identification.

Comparison of Protein Profiling and Inference Software

We compared IsoformResolver to other programs used for protein inference (IDPicker, Panoramics, Phenyx, Scaffold, TPP ProteinProphet). The programs varied with respect to input/output format, ease of use, and other features (summarized in Suppl. Table S2, Supporting Information). Here, we focused on their differences with respect to protein inference, protein grouping, how they dealt with indistinguishable proteins, their ability to handle large data sets, and comparison of results between different data sets.

Protein Inference. We first compared software with respect to protein inference on a single LC−MS/MS run (Suppl. Table S1, Data set 1D, Supporting Information). The numbers of peptides and proteins reported by each program were comparable and default parameters were used in each case, with settings chosen to yield comparable numbers of identified peptides. One complication was that Phenyx, Scaffold, and ProteinProphet integrate peptide identification algorithms into the software, each using different underlying methods to choose peptides, assess false assignments, and evaluate low scoring MS/MS spectra. This introduced variations in identified peptides, which complicated the comparison of protein identifications. Therefore, IsoformResolver was used to specify ISD groups from the peptides identified by each program. In this way, we could assess proteins identified by each program that were within the same ISD group, allowing differences in protein inference rather than differences in peptides to be evaluated.

Each program yielded proteins corresponding to 255−295 ISD groups. Of the 238 groups common to all six programs, 60 contained proteins that were distinct and unambiguously identified by all. In order to minimize differences due to peptide variations, 112 of the remaining 178 ISD groups were selected because all programs mapped identical proteins to the peptides in these groups (termed “meta-peptides” by ref 22). We inspected and compared proteins inferred for these 112 ISD groups.

Certain programs showed greater similarities in their protein identifications. Programs that reported all indistinguishable proteins as primary (TPP ProteinProphet, Panoramics, Scaffold, and IsoformResolver in its default mode) showed greater similarities in protein identification with each other, compared to programs which selected a single, representative protein from among each indistinguishable set (IsoformResolver in its representative protein selecting mode, Phenyx, and IDPicker). We identified five different cases. In 20 of 112 ISD groups (Case 1), the same proteins were identified by all 6 programs (Figure 9, see Suppl. Table S3 and Suppl. Worksheet:3.xlsx for the entire analysis, Supporting Information). In 34 ISD groups (Case 2), identical proteins were inferred by programs which selected and displayed all primary proteins, and by programs which displayed only representative proteins, although the proteins differed between the two program types. In 53 ISD groups (Case 3), proteins were identical among programs that displayed primary proteins, but nonidentical among programs that selected representative proteins. In 3 ISD groups (Case 4), the proteins were nonidentical among programs displaying primary proteins but identical among those selecting representative proteins. The remaining 2 ISD groups (Case 5) showed no agreement in proteins identified between the two kinds of programs. Thus, agreement was generally found between programs that selected primary proteins, while programs that selected representative proteins often disagreed with each other, and sometimes chose proteins that none of the other programs inferred. Similarly, analysis of the Sigma-Aldrich UPS1 sample of purified human proteins, where true and false protein identifications could be determined, showed that programs reporting primary proteins yielded more true assignments than programs reporting representative proteins (data not shown). We conclude that reporting primary proteins yields greater agreement after protein inference, whereas representative proteins, while convenient for simplifying output, loses important information.

Protein Display. An important difference between these programs was how they displayed bridge peptides. Phenyx, Panoramics, Scaffold, and ProteinProphet replicated bridge peptides, listing them redundantly with proteins that shared them. IDPicker dealt with bridge peptides by assigning them to only one protein and discarding them from others. When ProteinProphet and IsoformResolver profiles were compared (Suppl. Table S1, Data set 1C, Supporting Information), 777 MSD groups were found in common by both programs. ProteinProphet displayed secondary (subset) proteins within its protein groups (e.g., as in Suppl. Figure S3a, Supporting Information), but separated protein groups that shared bridge peptides. By contrast, IsoformResolver listed each peptide together within their MSD group, therefore bridge peptides were neither overrepresented nor underrepresented (Suppl. Figure S3b, Supporting Information), and reported the MSD groups adjacently in the output. By separating protein groups that shared bridge peptides, 17 of the 777 MSD groups in IsoformResolver were displayed as 34 protein groups in ProteinProphet, where members of each pair of related protein groups were separated far from each other in the output. This illustrates the advantage of a display that positions related proteins adjacently, in a manner that avoids peptide replication and redundancy.

Compare Profiles. Finally we examined the ability of each program to compare results from two or more data sets. IsoformResolver, Scaffold, Phenyx, and IDPicker were each able to display differences between multiple data sets within a single protein inference profile. Scaffold and Phenyx only allowed comparison of individual LC−MS/MS runs, while IsoformResolver and IDPicker allowed for any number of LC−MS/MS data sets (Suppl. Table S2, Supporting Information).

We also examined the ability of each program to compare results from separate protein inference analyses, for example, data sets analyzed at different times and then compared retrospectively. All programs allowed primary proteins to be manually compared between separate analyses; however, differences in protein inference and shortcomings of protein reporting led to overestimates of variation between analyses. This was alleviated by reporting protein groups, as allowed by IsoformResolver and IDPicker. However, IDPicker identified protein groups sequentially per profile, preventing their comparison against protein groups from other protein profiles. Only IsoformResolver had a stable (ISD) numbering scheme that allowed uniform comparisons between different experiments.

■ CONCLUSIONS

In this study, we describe IsoformResolver in detail for the first time. We demonstrate that protein inference exacerbates volatility in protein identifications, such that small changes in peptides lead to greater changes in the inferred proteins. We show that protein inference causes significant protein variation introduced by LC−MS/MS sampling in technical replicates, and even when peptides are completely overlapping between full data sets and simulated subsets. When many data sets are compared, protein repeatability is improved by pooling data sets at the peptide level and performing inference once, instead of performing inference on each experiment and aggregating the results. However, the pooled analysis loses important information gained by analyzing each experiment individually.

Underlying the problem of protein volatility is the question of how to select between indistinguishable proteins, inferred as present but not distinguishable from other equally possible candidates. Indistinguishable proteins must be counted singly and yet must be linked to multiple protein identifiers, because reporting all proteins in an indistinguishable set overestimates their presence, but reporting only one of several proteins loses valuable information. No single method of reporting protein identifiers—listing all proteins, primary proteins, concatenated identifiers, or representative proteins—completely solves the problem of underrepresenting or overrepresenting proteins in the sample due to protein inference.

Another important question is how to treat peptides that bridge multiple primary proteins. The results can be misleading when protein inference programs either assign the bridge peptides to only one protein arbitrarily, or else replicate the peptides and match them redundantly to multiple proteins, which may underestimate or overestimate the peptide evidence for a protein. We find that in complex protein databases like the human proteome, the number of bridge peptides increases as more peptides are identified with higher depth (e.g., see Data set 1 in Suppl. Table S1, Supporting Information).

IsoformResolver addresses all of these problems by reporting proteins and peptides in the context of MSD and ISD groups, developed using a peptide-centric strategy which lists each peptide once, and matching each observed peptide to all proteins that share its sequence. In this way, primary, secondary, and indistinguishable proteins can be immediately assessed by the presence or absence of distinguishing peptides, and are clearly marked in the output. Displaying proteins in the context of MSD groups avoids the problems of listing peptides redundantly or arbitrarily assigning them to one primary protein. Displaying primary, indistinguishable, and secondary proteins adjacently avoids loss of information about their relatedness, and allows the experimentalist, not the software, to decide which proteins are most likely present.

By displaying MSD groups adjacently and linked by ISD groups, all proteins linked by shared peptides can be listed together, even when the peptides are not observed experimentally. We show that proteins within ISD groups are usually derived from the same gene or products of gene duplication, exhibiting functional relationships which reflect their underlying sequence identity. Importantly, experimentally observed peptides and proteins can be mapped to protein identifiers which are invariant for a given database, lending stability to the protein profiles by allowing comparisons to be made between experiments analyzed at different times and using different software. ISD groups also allow IsoformResolver to facilitate comparison between data sets by spectral counting, by allowing related proteins to be listed adjacently.

In summary, protein inference remains a challenging problem, but the approach used by IsoformResolver, of converting a protein database into a peptide-centric format in which all nonredundant peptides are premapped to proteins, and all proteins are mapped to ISD groups, helps counteract many ambiguities introduced by the inference problem. In addition, when large data sets are involved, or many data sets must be compared, the algorithms employed by IsoformResolver allow greatly increased speed in execution time compared to other software. Presenting protein and peptide results in the context of MSD and ISD groups is a logical, complete, and concise way to display proteomics information, which solves problems in comparing data sets of high complexity from shotgun proteomics.

Software and peptide-centric database files are available upon request.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Supplemental figures and table. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Natalie G. Ahn, Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309-0215. Phone: 303-492-4799. Fax: 303-492-2439. E-mail: natalie.ahn@colorado.edu.

Notes

[‖] Deceased January 8, 2009

Present Addresses

[§] Brian Eichelberger, Dept. of Chemistry, John Brown University, Siloam Springs, AR 72761

■ ACKNOWLEDGMENT

We are indebted to Chia-Yu Yen, John Prince, Brian Searle, David Tabb, and Richard Johnson for valuable discussions. Funding was provided by an NRSA Molecular Biophysics Predoctoral Training Grant GM065103 (K.M.-A.), NIH R01 CA118972 (N.G.A.), and NIH R01 CA126240 (K.A.R.).

■ ABBREVIATIONS

ISD, *in silico*-derived protein groups; MSD, MS/MS-derived protein groups; DTA, file format for MS/MS spectra; MS/MS, mass spectrum; RP, reversed-phase; MGF, concatenated DTA file; FDR, false discovery rate; FP, false positive; TP, true positive; GO, Gene Ontology database of gene classifications; IPI, International Protein Index; SAX, strong anion exchange; TPP, Trans-Proteomic Pipeline.

■ REFERENCES

(1) Rappsilber, J.; Mann, M. What does it mean to identify a protein in proteomics? *Trends Biochem Sci.* **2002**, *27*, 74–78.

(2) Clamp, M.; Fry, B.; Kamal, M.; Xie, X.; Cuff, J.; Lin, M.; Kellis, M.; Lindblad-Toh, K.; Lander, E. Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 19428–19433.

(3) Kersey, P.; Duarte, J.; Williams, A.; Karavidopoulou, Y.; Birney, E.; Apweiler, R. The International Protein Index, An integrated database for proteomics experiments. *Proteomics* **2004**, *4*, 1985–1988.

(4) Nesvizhskii, A.; Aebersold, R. Interpretation of shotgun proteomics data: the protein inference problem. *Mol. Cell. Proteomics* **2005**, *4*, 1419–1440.

(5) Tabb, D.; McDonald, W.; Yates, J. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **2002**, *1*, 21–26.

(6) Nesvizhskii, A.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, 4646–4658.

(7) Yang, X.; Dondeti, V.; Dezube, R.; Maynard, D.; Geer, L.; Epstein, J.; Chen, X.; Markey, S.; Kowalak, J. DBParser: Web-based software for shotgun proteomics data analyses. *J. Proteome Res.* **2004**, *3*, 1002–1008.

(8) Allet, N.; Barrillat, N.; Baussant, T.; Boiteau, C.; Botti, P.; Bougueleret, L.; Budin, N.; Canet, D.; Carraud, S.; Chiappe, D.; Christmann, N.; Colinge, J.; Cusin, I.; Dafflon, N.; Depresle, B.; Fasso, I.; Frauchiger, P.; Gaertner, H.; Gleizes, A.; Gonzalez-Couto, E.; Jeandenans, C.; Karmime, A.; Kowall, T.; Lagache, S.; Mahé, E.; Masselot, A.; Mattou, H.; Moniatte, M.; Niknejad, A.; Paolini, M.; Perret, F.; Pinaud, N.; Ranno, F.; Raimondi, S; Reffas, S.; Regamey, P. O.; Rey, P. A.; Rodriguez-Tomé, P.; Rose, K.; Rossellat, G.; Saudrais, C.; Schmidt, C.; Villain, M.; Zwahlen, C. *In vitro* and *in silico* processes to identify differentially expressed proteins. *Proteomics* **2004**, *4*, 2333–51.

(9) Resing, K.; Meyer-Arendt, K.; Mendoza, A.; Aveline-Wolf, L.; Jonscher, K.; Pierce, K.; Old, W.; Cheung, H.; Russell, S.; Wattawa, J.; Goehle, G.; Knight, R.; Ahn, N. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* **2004**, *76*, 3556–3568.

(10) Old, M.; Shabb, J.; Houel, S.; Wang, H.; Couts, K.; Yen, C.; Litman, E.; Croy, C.; Meyer-Arendt, K.; Miranda, J.; Brown, R.; Witze, E.; Schweppe, R.; Resing, K.; Ahn, N. Functional proteomics identifies targets of phosphorylation by B-Raf signaling in melanoma. *Mol. Cell* **2009**, *34*, 115–131.

(11) Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N.; Old, W. Quantifying the impact of chimera MS/MS spectra on peptide identification in large scale proteomics studies. *J. Proteome Res.* **2010**, *9* (8), 4152–4160.

(12) Eng, J.; McCormack, A.; Yates, J. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.

(13) Perkins, D.; Pappin, D.; Creasy, D.; Cottrell, J. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–3567.

(14) Elias, J.; Gygi, S. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4*, 207–214.

(15) Blanco, L.; Mead, J. A.; Bessant, C. Comparison of novel decoy database designs for optimizing protein identification searches using ABRF sPRG2006 standard MS/MS data sets. *J. Proteome Res.* **2009**, *8*, 1782–1791.

(16) Yen, C.; Russell, S.; Mendoza, A.; Meyer-Arendt, K.; Sun, S.; Cios, K.; Ahn, N.; Resing, K. Improving sensitivity in shotgun proteomics using a peptide-centric database with reduced complexity: Protease cleavage and SCX elution rules from data mining of MS/MS spectra. *Anal. Chem.* **2006**, *78*, 1071–1084.

(17) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **2004**, *76*, 3908–3922.

(18) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal. Chem.* **2005**, *77*, 6364–6373.

(19) Sun, S.; Brown, R.; Yen, C.; Meyer-Arendt, K.; Old, W.; Pierce, K.; Ahn, N.; Cios, K.; Resing, K. Improved validation of peptide MS/MS assignments using spectral intensity prediction and full annotation of fragment ions. *Mol. Cell. Proteomics* **2007**, *6*, 1–17.

(20) Searle, B.; Turner, M.; Nesvizhskii, A. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J. Proteome Res.* **2008**, *7*, 245–253.

(21) Feng, J.; Naiman, D.; Cooper, B. Probability model for assessing proteins assembled from peptide sequence inferred from tandem mass spectrometry data. *Anal. Chem.* **2007**, *79*, 3901–3911.

(22) Zhang, B.; Chambers, M.; Tabb, D. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.* **2007**, *6*, 3549–3557.

(23) Ma, Z.; Dasari, S.; Chambers, M.; Litton, M.; Sobecki, S.; Zimmerman, L.; Halvey, P.; Schilling, B.; Drake, P.; Gibson, B.; Tabb, D. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.* **2009**, *8*, 3872–3881.

(24) Colinge, J.; Masselot, A.; Cusin, I.; Mahe, E; Niknejad, A; Argoud-Puy, G.; Reffas, S.; Bederr, N.; Gleizes, A.; Rey, P.; Bougueleret, L. High-performance peptide identification by tandem mass spectrometry allows reliable automatic data processing in proteomics. *Proteomics* **2004**, *4*, 1977–1984.

(25) Bradshaw, R.; Burlingame, A.; Carr, S.; Aebersold, R. Reporting protein identification data: the next generation of guidelines. *Mol. Cell. Proteomics* **2006**, *5*, 787–788.

(26) Taylor, C. F.; Paton, N. W.; Lilley, K. S.; Binz, P. A.; Julian, R. K., Jr; Jones, A. R.; Zhu, W.; Apweiler, R.; Aebersold, R.; Deutsch, E. W.; Dunn, M. J.; Heck, A. J.; Leitner, A.; Macht, M.; Mann, M.; Martens, L.; Neubert, T. A.; Patterson, S. D.; Ping, P.; Seymour, S. L.; Souda, P.; Tsugita, A.; Vandekerckhove, J.; Vondriska, T. M.; Whitelegge, J. P.; Wilkins, M. R.; Xenarios, I.; Yates, J. R., 3rd; Hermjakob, H. The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* **2007**, *25*, 887–893.

(27) Old, W.; Meyer-Arendt, K.; Aveline-Wolf, L.; Pierce, K.; Mendoza, A.; Sevinsky, J.; Resing, K.; Ahn, N. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* **2005**, *4*, 1487–1502.

(28) Hulsen, T.; de Vlieg, J.; Alkema, W. BioVenn — a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics* **2008**, *9*, 488–493.

(29) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–1372.

(30) Zhang, Y.; Wen, Z.; Washburn, M.; Florens, L. Refinements to label free proteome quantitation: How to deal with peptides shared by multiple proteins. *Anal. Chem.* **2010**, *82*, 2272–2281.

(31) Jin, S.; Daly, D.; Springer, D.; Miller, J. The effects of shared peptides on protein quantitation in label-free proteomics by LC—MS/MS. *J. Proteome Res.* **2008**, *7*, 164–169.