Supplementary Information for: “Improved Protein Inference from Multiple Protease Bottom-Up Mass Spectrometry Data”

Rachel M. Miller[1], Robert J. Millikin[1], Connor V. Hoffmann[1†], Stefan K. Solntsev[1†], Gloria M. Sheynkman[1†], Michael R. Shortreed[1] and Lloyd M. Smith[1*]

[1] Department of Chemistry, University of Wisconsin, 1101 University Avenue, Madison, Wisconsin 53706, United States

Table of Contents

Supplementary Experimental Methods:

Cell Culture. The Jurkat cell line (TIB-152) chosen for this study was obtained from the American Type Culture Collection (ATCC, Manassas, VA). Cells were cultured in 10% Fetal Bovine Serum (FBS) and 90% RMPI medium at 37 °C and grown to a concentration of approximately 1.3 x 10^6 cells/mL. Six cell aliquots of approximately 3.2 x 10^7 cells each were centrifuged at 180 x g and 4 °C for 10 minutes. The cell pellets were washed twice with ice-cold PBS buffer. The final cell pellets were flash frozen and stored at -80 °C until needed.

Protein Extraction. Flash frozen Jurkat cell pellets were thawed on ice. Cells were lysed by pipetting the pellet repeatedly with SDT lysis buffer ([4% SDS, 500 mM Tris-HCl (pH 7.4)] and 180 mM dithiothreitol (DTT), added in a 5:1 volume ratio to that of the cell pellet) followed by a 5-minute incubation at 95 °C. Lysate was probe sonicated on ice for 3 to 5 minutes, cycling between 30 seconds sonication and 30 seconds rest.

Filter-Aided Sample Preparation. Approximately 150 µg of protein from each aliquot of lysate was transferred to a 100K Amicon Ultra filter (Millipore, Billerica, MA). A modified FASP protocol was utilized to accommodate differing protease digestion conditions. The original FASP protocol was employed until the final wash with ammonium bicarbonate (pH 7.8)[12]; then each filter was washed two additional times with the buffer system appropriate for its specific protease. Protease aliquots were then added to their respective filters. Optimized digestion conditions for each protease are given in Table S1. Following digestion, each filter was centrifuged at

14,000xg for 15 minutes to recover digested peptides. The amount of peptide recovered was quantified via Pierce BCA assay (ThermoFisher Scientific).

Peptide Fractionation. At least 100 µg of each peptide digest was fractionated at high pH on a Shimadzu HPLC using a Phenomenex C18 Gemini 3 µ, 110Å, 3.0 x 150 mm column. The buffers used for separation were 20 mM ammonium formate (pH 10) in water (mobile phase A, MA), 20 mM ammonium formate (pH 10) in 70% acetonitrile (mobile phase B, MB). The flow rate was 0.5 mL/min and the binary gradient was: 0% MB for 15 minutes, linear ramp to 100% MB over 45 minutes, hold at 100% MB for 5 minutes, linear descent to 0% B over 2 minutes followed by equilibration at 0% MB for 20 minutes. Eleven 1 mL fractions of peptides were collected for all proteases with the exception of the tryptic digest where only 10 fractions were obtained. Fractions were lyophilized via SpeedVac and stored at -80 °C.

LC-MS/MS Analysis. Each lyophilized fraction was reconstituted in 5% acetonitrile and 1% formic acid, followed by chromatography on a nanoACQUITY LC system (Waters, Milford, MA) interfaced with a Thermo Scientific LTQ Orbitrap Velos mass spectrometer (Thermo Fisher, Waltham, MA) using a 20 cm reverse-phase capillary column packed with 3 µm C18 beads. Buffers used were 0.2% formic acid in water (mobile phase A, MA) and 0.2% formic acid in acetonitrile (mobile phase B, MB). Full scans from 300-1500 *m/z* were collected at a resolution of 60,000. These MS1 scans were followed by top 10 precursor HCD fragmentation to produce spectra at a resolution of 7,500. Precursor fragmentation repeat count was set to two, and dynamic exclusion was set to 60 s.

Table S1: Protease Specific Digestion Conditions

Protease	Protein:Enzyme Ratio	Buffer System	Temperature of Digest (°C)	Length of Digest (hours)
Arg-C	100:1	Incubation Buffer: 50 mM Tris-HCl (pH 7.6), 5 mM $CaCl_2$, and 2 mM EDTA Activation Buffer: 50 mM Tris-HCl (pH 7.6), 50 mM DTT, and 2 mM EDTA Combined Incubation and Activation buffer in 9:1 ratio	37	16
Asp-N	100:1	50 mM sodium phosphate (pH 8.0)	25	16
Chymotrypsin	100:1	100 mM Tris-HCl (pH 8.0) and 10 mM $CaCl_2$	25	12
Glu-C	100:1	25 mM Ammonium Bicarbonate (pH 7.8)	25	16
Lys-C	100:1	25 mM Tris-HCl (pH 8.5), 1 mM EDTA, and 4 M urea	37	16
Trypsin	50:1	50 mM Ammonium Bicarbonate (pH 7.8)	37	16

Table S2: MetaMorpheus Search Parameters

Global Search Parameters
Search Mode
Classic Search
In-silico Digestion Parameters
Generate target proteins Generate decoy proteins Generate reversed decoys Max Missed Cleavages :2 Initiator Methionine: Variable Max Modification Isoforms: 1024 Min Peptide Length: 7 (5 for ProteinProphet Comparison) Max Peptide Length: none Max mods per peptide: 2
Fragment Ion Search Parameters
Dissociation Type: HCD Max Threads: 39 Max Fragment Mass (Da): 30000 N-Terminal Ions C-Terminal Ions
Mass Difference Acceptors
1 Missed Monoisotopic Peak
Ambiguity Parameters
Report PSM ambiguity
Scoring Options
Minimum score allowed: 5
Post-Search Analysis
Apply protein parsimony and construct protein groups

Table S3: Modifications for G-PTM-D

G-PTM-D Modifications		
Common Biological		
Acetylation on K Acetylation on X (Prot N-Term) ADP-ribosylation on S Butyrylation on K Carboxylation on D Carboxylation on E Carboxylation on K Citrullination on R Crotonylation on K Dimethylation on K Dimethylation on R	Formylation on K Glu to PyroGlu on Q (Prot N-Term) Glutarylation on K HexNAc on Nxs HexNAc on Nxt HexNAc on S HexNAc on T Hydroxybutyrylation on K Hydroxylation on K Hydroxylation on N Hydroxylation on P	Malonylation on K Methylation on K Methylation on R Nitrosylation on C Nitrosylation on Y Phosphorylation on S Phosphorylation on T Pyridoxal phosphate on K Succinylation on K Sulfonation on Y Trimethylation on K
Common Artifact		
Ammonia loss on C (Pep N-Term) Ammonia loss on N Carbamyl on C Carbamyl on K	Carbamyl on M Carbamyl on R Carbamyl on X (Pep N-Term)	Deamidation on N Deamidation on Q Water Loss on E (Pep N-Term)
Metal		
Calcium on D Calcium on E Cu[I] on D Cu[I] on E Fe[II] on D Fe[II] on E	Fe[III] on D Fe[III] on E Magnesium on D Magnesium on E Potassium on D	Potassium on E Sodium on D Sodium on E Zinc on D Zinc on E

Table S4: Comet Search Parameters

Search Parameters		
Search		
Decoy search:1	Peff format:1	Num threads: -1
Masses		
Peptide mass tolerance: 20.00 Peptide mass units: 2	Mass type parent: 1 Mass type fragment: 1	Precursor tolerance type: 1 Isotope error: 1
Search Enzyme		
Search enzyme number: (1- Trypsin, 3- Lys-C, 5- Arg-C, 6- Asp-N, 8- Glu-C and 10- Chymotrypsin) Num enzyme termini:2 2 Allowed missed cleavage: 2		
Modifications		
Variable mod 01: 15.9949 M 0 2 -1 0 0	Max variable mods in peptide: 5	Require variable mod: 0
Fragment Ions		
Fragment bin tol: 0.4 Fragment bin offset: 0.4 Theoretical fragment ions: 1 Use A ions: 0	Use B ions: 1 Use C ions: 0 Use X ions: 0	Use Y ions: 1 Use Z ions: 0 Use NL ions: 0
Misc. Parameters		
Digest mass range: 600.0 5000.0 Num results: 1000 Skip researching: 1 Max fragment charge: 3	Max precursor charge: 6 Nucleotide reading frame: 0 Clip nterm methionine: 0	Spectrum batch size: 0 Decoy prefix: DECOY Equals I and L: 1
Spectral Processing		
Minimum peaks: 10 Minimum intensity: 0	Remove precursor peak: 0 Remove precursor tolerance: 1.5	Clear mz range: 0.0 0.0

Additional Modifications		
Add Cterm peptide: 0.0	Add L leucine: 0.0000	Add U selenocysteine: 0.0000
Add Nterm peptide: 0.0	Add I isoleucine: 0.0000	Add R arginine: 0.0000
Add Cterm protein: 0.0	Add N asparagine: 0.0000	Add Y tyrosine: 0.0000
Add Nterm protein: 0.0	Add D aspartic acid: 0.0000	Add W tryptophan: 0.0000
Add G glycine: 0.0000	Add Q glutamine: 0.0000	Add B user amino acid: 0.0000
Add A alanine: 0.0000	Add K lysine: 0.0000	Add J user amino acid: 0.0000
Add S serine: 0.0000	Add E glutamic acid: 0.0000	Add X user amino acid: 0.0000
Add P proline: 0.0000	Add M methionine: 0.0000	Add Z user amino acid: 0.0000
Add V valine: 0.0000	Add O ornithine: 0.0000	
Add T threonine: 0.0000	Add H histidine: 0.0000	
Add C cysteine: 57.021464	Add F phenylalanine: 0.0000	

Table S5: TPP- ProteinProphet Parameters

Input is from iProphet	false
Import XPRESS protein ratios	false
Import ASPARatio protein ratios and pvalues	false
Import Libra protein ratios	false
Do not include zero probability protein entries in output	true
Do not report protein length	false
Report(calculated) protein molecular weight	false
Icat data	false
N-glycosylation data	false
Delude (do not look up ALL proteins corresponding to shared peps)	false
Do not use Occam's razor for shared peps	false
Do not assemble protein groups	false
Normalize NSP using protein length	false
Use expected number of ion instance to adjust the peptide probabilities prior to NSP adjustment	false
Check peptide's Protein Weight against the threshold	false

Table S6: ProLuCID Search Parameters

Search Mode	
Primary score type	1-XCorr
Secondary score type	2-zScore
Locus type	0-accession
Charge disambiguation	0
Atomic enrichment	0-no labeling
Min match	5
Peak rank threshold	200
Candidate peptide threshold	500
Num output	5
Is decharged	0
Fragmentation method	CID
Pre process	1-do XCorr like preprocessing
Isotopes	
Precursor	Mono
Fragment	Mono
Num peaks	0
Tolerance	
Precursor high	4500
Precursor low	4500
Precursor mass accuracy	5
Fragment ion mass accuracy	20
Precursor mass limits	
Minimum	600
Maximum	1600
Peptide length limits	
Minimum	7

Num peak limits	
Minimum	25
Maximum	5000
Max num diff mods	0
Modifications	
Static Mods	C & U mass shift: 57.02146
Diff Mods	M mass shift: 15.9949146
Enzyme Info	
Specificity	2-both ends
Max num internal mis cleavage	2
Name, type and residues	Depends on protease

Table S7: DTASelect2 Parameters

Enzyme number	0
Diff search options	16 M
Include peptides regardless of cleavage status	-y0
Peptide FDR	--fp 0.01
Protein FDR	--pfp 0.01
Decoy identifier	--decoy DECOY_

Supplementary Methods for Data Analysis:

The implementation of the “integrated” multi-protease protein inference algorithm in MetaMorpheus required that peptide confidence (*q*-values) be calculated separately for each protease, and that the peptide identification be associated with their protease of origin. These two modifications to MetaMorpheus’ conventional protein inference are explained below.

Peptide q-Values. MetaMorpheus uses q-values as an assessment of confidence in a given identification, describing the minimum FDR threshold at which the peptide would exist within thedataset. Peptide spectral match (PSM) q-values are calculated by ranking the identifications by score, then calculating the ratio of cumulative decoy to targets identifications for each PSM. The length of the peptide is directly correlated to the number of fragment matching opportunities a target or decoy spectra has to the theoretical database, and each protease produces peptides with differing length distributions. In MetaMorpheus, the number of fragment ion matches is used to determine the score of the PSM, therefore target and decoy score distributions differ by protease (Figure S1). PSM scores are subsequently used for ranking prior to q-value calculation. High scoring decoys from one protease should not penalize the q-value of target PSMs from another protease, hence the need to calculate peptide confidence levels separately for each protease in order to maintain the integrity of the value.

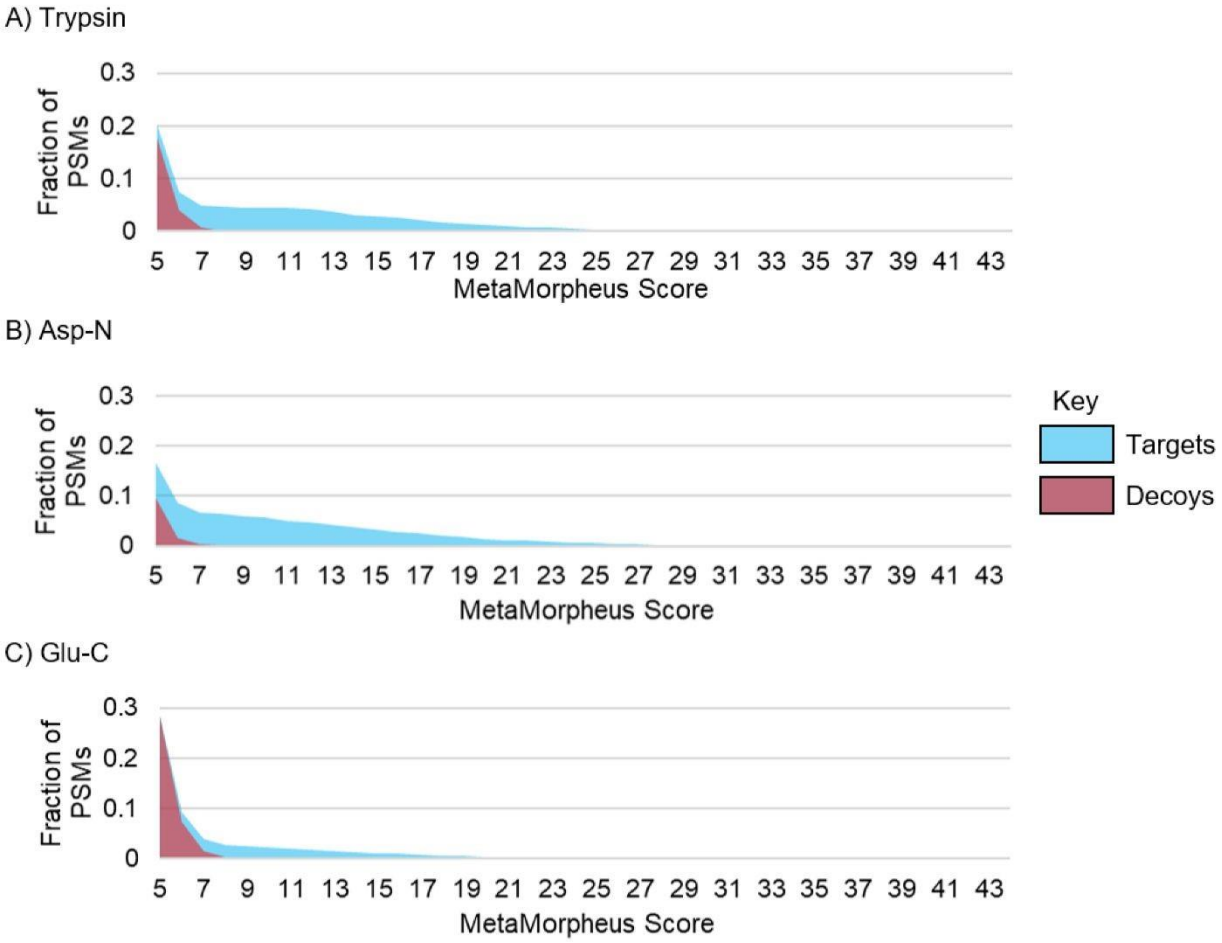

Figure S1: Distribution of target and decoy PSM scores resulting from digestion with A) Trypsin, B) Asp-N and C) Glu-C.

Associating Peptides with Their Protease of Origin.

Ignoring the sequence-protease relationship can add unnecessary ambiguity to multi-protease protein inference results. The integrated multi-protease protein inference algorithm maintains this relationship, enabling accurate determination of whether a peptide should be classified as shared or unique, and which proteins the peptide could have potentially originated. A sequence of amino acids could be unique within the proteome for digestion with one protease but shared for another, or be shared among different proteins based on its originating protease (Figure S2). These situations arise for 5.3% and 12% of the 42,419 proteins present in the UniProt Human Canonical and Isoform reviewed database, respectively. For example, if the peptide "FHSMASR", from Figure S2, is identified from spectra resulting from Lys-C digestion, it could have resulted from the digestion of Q86UV7 or Q86UV6. If the protease that resulted in the production of this peptide is unknown, then the number of possible parent proteins increases from 2 to 5 (Q86T4 or Q86XT4-2 or Q86UV6 or Q86UV6-2 or Q86UV7).

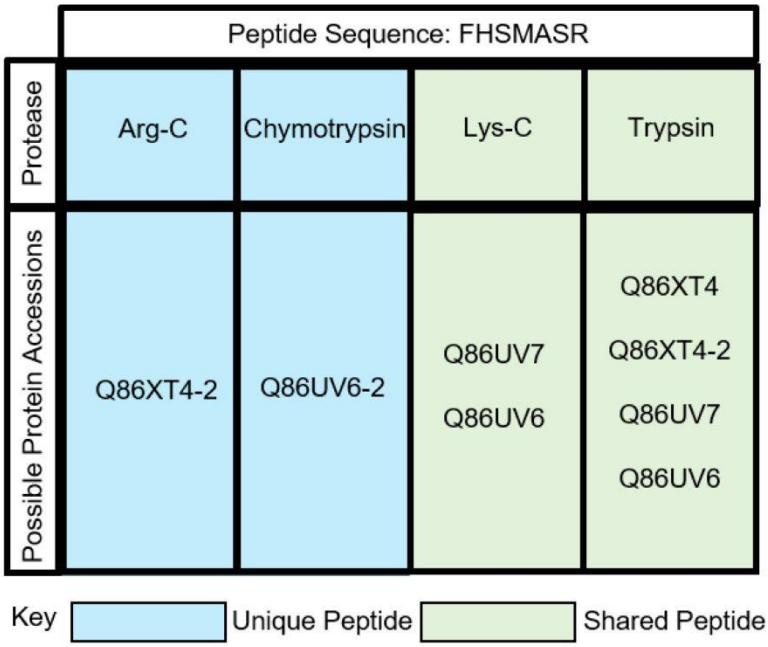

Figure S2: Depiction of how the peptide sequence "FHSMASR" can be attributed to different protein groups based on which protease it originated from. Sections in blue indicate that the peptide sequence is unique to a single protein accession for the assigned protease whereas sections in green indicate the same peptide sequence is shared for the assigned protease. This shows how the association of a peptide sequence with its protease of origin can eliminate unnecessary protein group ambiguity.

Supplementary Information for MetaMorpheus' Separate and Integrated Multi-Protease Comparison:

The total number of protein groups identified with the separate multi-protease approach is greater than that of the integrated multi-protease approach at a 1% protein FDR threshold (Table 2). To ensure that this difference in the total number of protein groups does not bias the analysis of the accuracy of the multi-protease protein inference approaches, the results were compared at the total number of protein groups present at the 1% FDR threshold for both the integrated and separate multi-protease approaches (7,472 protein groups and 7,716 protein groups, respectively). The results of this analysis are summarized in Tables S8 and S9. A decrease in the number of *Arabidopsis thaliana* identifications and the corresponding false positive rate was observed for the integrated approach, further indicating that the integrated approach provides more accurate protein group results than the separate approach.

Table S8: Comparison of Entrapment Results of the Top 7,472 Protein Groups from the Separate and Integrated Multi-Protease Approaches.

	Separate Multi-Protease Approach	Integrated Multi-Protease Approach	Percent Change
Number of Human Protein Groups	7,253	7,255	+0.03%
Number of *Arabidopsis thaliana* Protein Groups	219	217	-0.93%
False Positive Rate	2.93%	2.90%	-1.02%

Table S9: Comparison of Entrapment Results of the Top 7,716 Protein Groups from the Separate and Integrated Multi-Protease Approaches.

	Separate Multi-Protease Approach	Integrated Multi-Protease Approach	Percent Change
Number of Human Protein Groups	7,400	7,407	+0.09%
Number of Arabidopsis thaliana Protein Groups	316	309	-2.22%
False Positive Rate	4.10%	4.00%	-2.44%

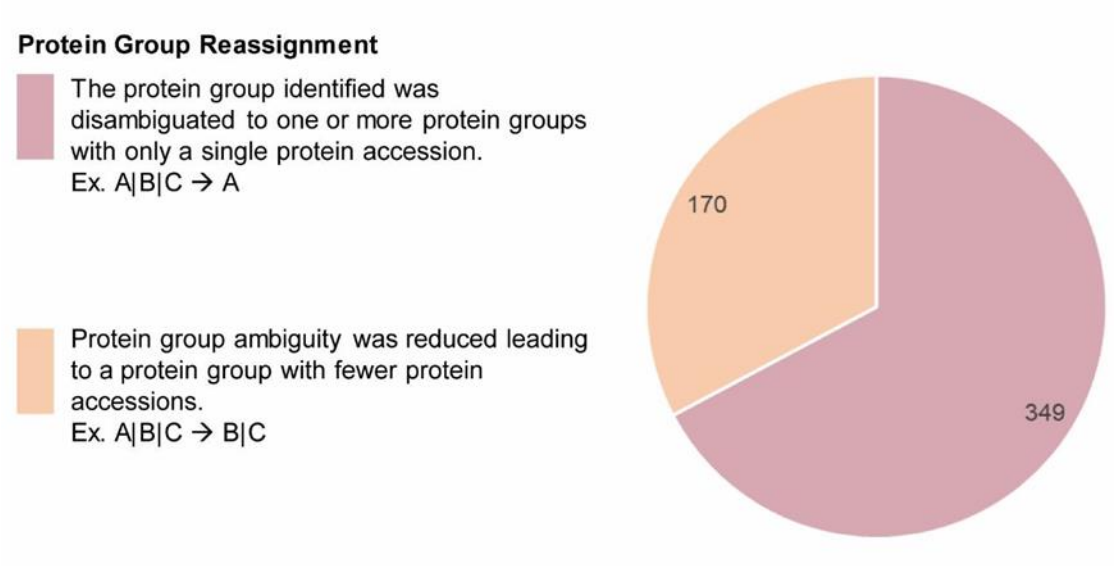

Figure S3: A majority of the protein groups unique to the separate approach (519 of 864) were disambiguated into simpler protein groups in the results of the integrated multi-protease approach and can be assigned to two distinct categories: disambiguation to one or more protein groups with a single protein member, or disambiguation to a protein group with fewer protein members. A pie chart representing the distribution of the 519 protein groups into

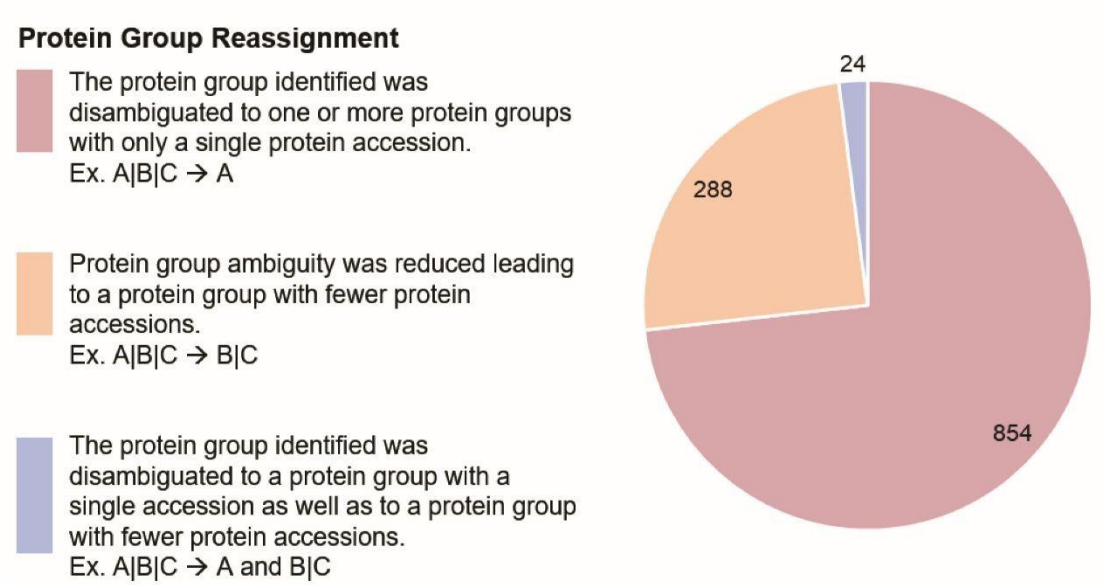

Figure S4: Almost all of the protein groups unique to the tryptic digest (1,166 of 1,202) were disambiguated into simpler protein groups in the results of the integrated multi-protease approach could be assigned to three distinct categories: disambiguation to one or more protein groups with a single protein member, disambiguation to a protein group with fewer protein members, or disambiguation to both a protein groups containing a single member

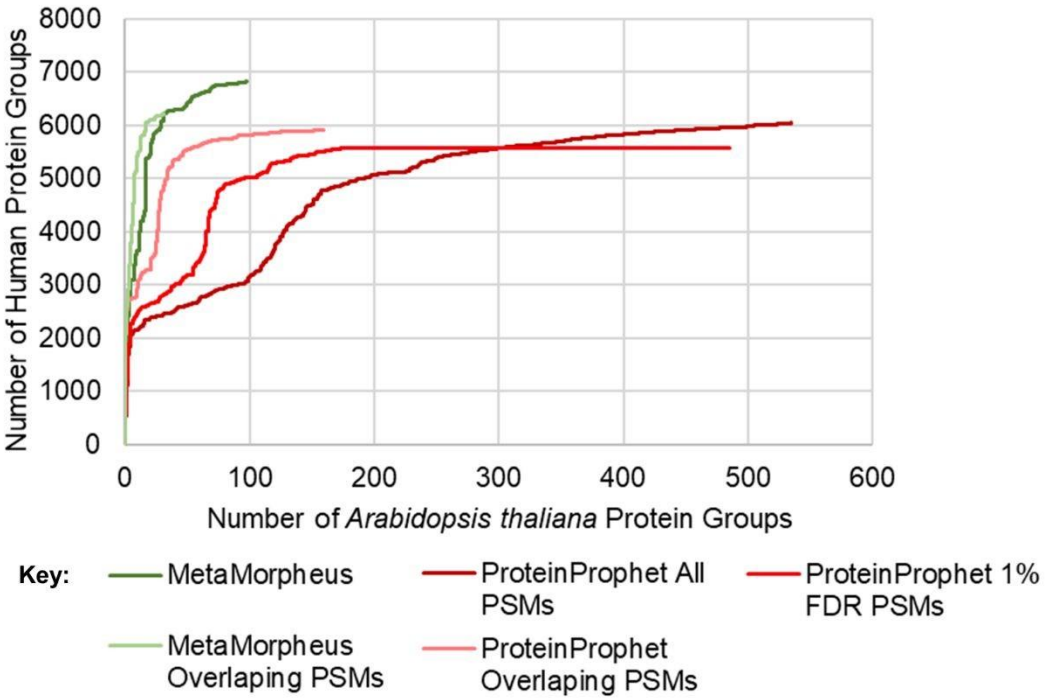

Figure S5: Curves comparing the ability of MetaMorpheus' and ProteinProphet's multi-protease protein inference algorithms to distinguish between human protein groups (true positives) and *Arabidopsis thaliana* protein groups (false positives) based on the PSMs used for protein inference. Limiting the PSMs used for protein inference to those that were identified in both MetaMorpheus and Comet searches provided increased accuracy for both algorithms compared to their un-filtered counterparts and provided a more unbiased comparison.

Table S10: Comparison of Peptide Sequences Identified by MetaMorpheus and Comet at 1% FDR.

	Comet	MetaMorpheus	Overlap
Number of Arg-C Peptide Identifications	11,148	17,270	10,109
Number of Asp-N Peptide Identifications	17,326	12,065	9,565
Number of Chymotrypsin Peptide Identifications	18,073	10,831	6,218
Number of Glu-C Peptide Identifications	14,828	10,956	8,666
Number of Lys-C Peptide Identifications	25,569	31,962	24,321
Number of Trypsin Peptide Identifications	24,873	29,497	23,058

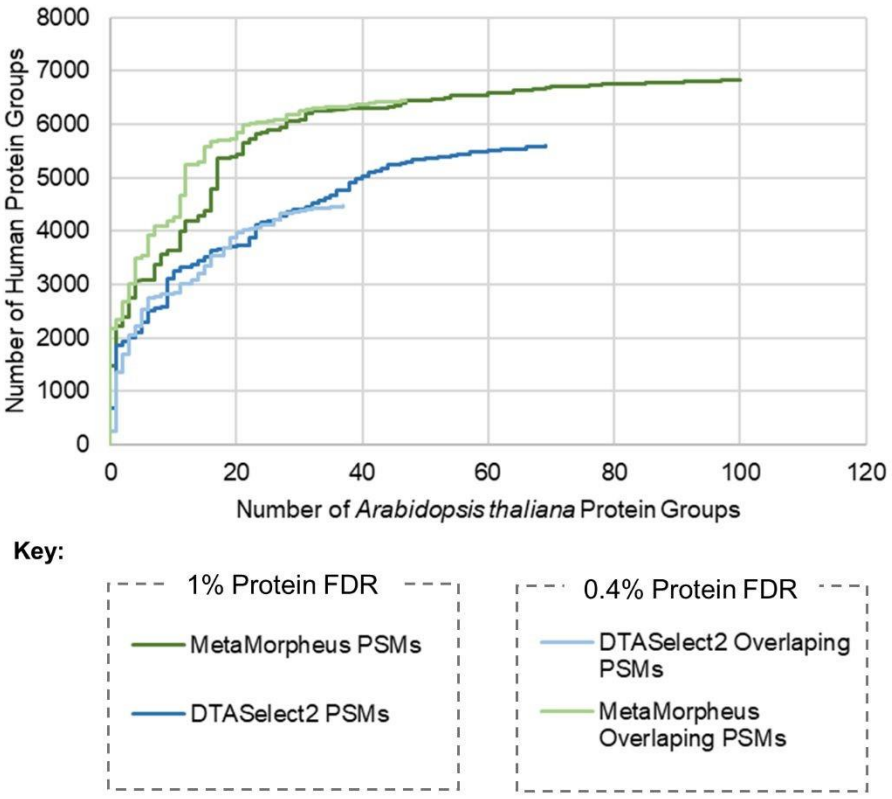

Figure S6: Curves comparing the ability of MetaMorpheus' and DTASelect2's multi-protease protein inference algorithms to distinguish between human protein groups (true positives) and *Arabidopsis thaliana* protein groups (false positives) based on the PSMs used for protein inference. Limiting the PSMs used for protein inference to those that were identified in both MetaMorpheus and ProLuCID searches provided increased accuracy for both algorithms compared to their un-filtered counterparts and provided a more unbiased comparison.

Table S11: Comparison of Peptide Sequences Identified by MetaMorpheus and ProLuCID at 1% FDR.

	ProLuCID	MetaMorpheus	Overlap
Number of Arg-C Peptide Identifications	16,696	17,064	13,122
Number of Asp-N Peptide Identifications	16,794	13,912	11,005
Number of Chymotrypsin Peptide Identifications	3,203	12,157	3,072
Number of Glu-C Peptide Identifications	9,629	13,423	8,316
Number of Lys-C Peptide Identifications	38,156	37,125	28,649
Number of Trypsin Peptide Identifications	26,440	41,935	26,137