Supplemental Materials

6-prot SpectrumMill Parameters

All MS data was analyzed using the Spectrum Mill software package v4.0 beta (Agilent Technologies, Santa Clara, CA). Spectral pre-processing filters used to exclude MS/MS spectra from further analysis included precursor charge >11, precursor MH+ > 10,000 Da, and poor quality as measured by not having a sequence tag length > 0 (i.e., minimum of two masses separated by the in-chain mass of an amino acid). MS/MS spectra were searched using Spectrum Mill for the purpose of training the MS/MS Charge Deconvolution algorithm described below that was developed for the SPS pipeline. The Spectrum Mill search algorithm has a similar onboard fragment charge assignment algorithm for high resolution spectra that uses mass spacing between peaks and a Chi2 metric between the intensities in an experimental isotope cluster and an averagine-derived theoretical isotope cluster (1). When the chi2 metric threshold is not met, a peak is treated as ambiguous charge and charge values of 1 to 3 are allowed for that peak during the matching phase of the search algorithm. Spectrum Mill searches were performed against a UniProt sequence database containing 28,189 entries that included the sequences of the 6 purified proteins present in the sample, common laboratory contaminants, proteolytic enzymes used, and the full proteome of Drosophila melanogaster (including isoforms and excluding fragments) downloaded from the UniProt web site on July 16, 2010 (2). Search parameters included: ESI Orbitrap or ESI Orbitrap HCD scoring parameters, enzyme specificity with maximum missed cleavages of 2 (trypsin, Arg-C, Glu-C, Asp-N), 3 (Lys-C), or 4 (chymotrypsin), 30% minimum matched peak intensity, +/- 10 ppm precursor mass tolerance, +/- 15 ppm product mass tolerance, and carbamidomethylation of cysteines as a fixed modification. Allowed variable modifications were oxidation of methionine, deamidation of asparagine, and pyro-glutamic acid modification at N-terminal glutamine with a precursor MH+ shift range of -18 to 64 Da. For the CNBr digest, no enzyme specificity was allowed, the additional fixed modification of methionine, homoserine lactone was included, and a precursor MH+ shift range of -50 to 5 Da was used. Identities interpreted for individual spectra were automatically designated as valid by optimizing score and delta rank1-rank2 score thresholds separately for each precursor charge state in each LC-MS/MS run while allowing a maximum false-discovery rate (FDR) of 1.2% at the spectrum level. That FDR threshold was implemented in target-decoy-fashion using reversed sequences of all proteins in the database. Since the sample contained no proteins from Drosophila melanogaster the FDR estimate could be checked and was found to have been underestimated. 2.4% (CID) and 2.7% (HCD) of the identified spectra were matched to fly peptides. CID and HCD identified spectra were then separated by precursor charge (2, 3, and 4 or greater) and again filtered by 1% spectrum-level FDR where any hit to a target or contaminate sequence was considered a true positive and any hit to a Drosophila sequence was considered a false positive.

MS/MS Charge Deconvolution

We performed *MS/MS Charge Deconvolution* on each spectrum with the purpose of identifying all isotopic envelopes and replacing each envelope with its monoisotopic peak of charge one. Deconvolution procedures for top-down MS/MS mass spectra, such as Thrash (3) and Xtract (4), generate a theoretical isotopic envelope for every candidate peak and charge and then compare it to the observed isotopic envelope. Our approach operates in the same manner except that it was designed and trained for bottom-up MS/MS spectra, where there is typically less evidence (ie less ^{13}C isotopes) supporting the presence of low charged (e.g. charge 2 and 3) ions than there is evidence for highly charged ions in top-down MS/MS spectra. For each candidate peak and charge state, a theoretical isotopic envelope was simulated assuming universal abundances of C, H, N, O, and S (as done in (3)). This envelope was then compared to the normalized isotopic envelope obtained from the spectrum by the Kullback-Leibler measure of divergence (5). Since only one or two peaks of each isotopic envelope are typically present in MS/MS spectra, isotopic envelopes of size 2 – 5 were considered. If the KL-divergence was below parameter β, the observed isotopic envelope was replaced with its monoisotopic peak of charge 1. Training of β was done by maximizing observed charge 1 b and y ion counts as a function of β in multiple identified spectra (Table 2).

Input: Peak mass tolerance k, divergence cutoff β, all expected isotopic envelopes E'_z for given charges z between 2 and 20, and MS/MS spectrum S with parent mass $\mathrm{PM}[S]$ and precursor charge $\mathrm{Z}[S] > 1$.

Output: $\mathrm{Deconv}(S, k, E', \beta)$ with all peaks converted to charge one.

```
$\mathrm{MonoIso}(p, z)$
    Return $(p * z) - z + \mathrm{mass}(\mathrm{H}^+)$

$\mathrm{ExtractEnv}(p, z, S, k)$
    $E \leftarrow [0,0,0,0,0]$
    For each $i$ from $0 \rightarrow 4$
        Add the cumulative intensity of all peaks within $k$ of $p + \frac{i}{z}$ to $E[i]$
    normalize $E$ (divide each value by the total intensity in $E$)
    Return $E$

$\mathrm{CompareEnv}(p, S, k, E')$
    $div_{min} \leftarrow \infty,\ E_{min} \leftarrow [\quad],\ z_{min} \leftarrow 0$
    For each charge $z$ from $\mathrm{Z}[S] \rightarrow 1$
        If $\mathrm{MonoIso}(p, z) > \mathrm{PM}[S]$ then continue
        $E \leftarrow \mathrm{ExtractEnv}(p, z, S, k)$
        If $\exists\ i$ st $E[i] > 12 * \mathrm{i}[p]$ then continue
        For each $j$ from $4 \rightarrow 1$
            $div \leftarrow \sum_{i=0}^{j} \left(E[i] * \log(E[i]/E'_z[i])\right)$
            If $div < div_{min}$ then $div_{min} \leftarrow div,\ E_{min} \leftarrow E[0 \ldots j]$, and $z_{min} \leftarrow z$
    Return $div_{min},\ E_{min}\ ,\ z_{min}$

$\mathrm{Deconv}(S, k, E', \beta)$
    $S_D \leftarrow S$
    For each $p \in S$ in increasing value
        $div_{min},\ E_{min},\ z_{min} = \mathrm{CompareEnv}(p, S, k, E')$
        If $div_{min} < \beta$
            Remove all peaks considered in $E_{min}$ from $S_D$ and $S$
            Add $p'$ to $S_D$ with $\mathrm{m}[p'] = \mathrm{MonoIso}(p, z_{min}),\ \mathrm{z}[p'] = 1,$ and $\mathrm{i}[p'] = \sum_{i=0} E_{min}[i]$
    Return $S_D$
```

Table S-1

	Precursor Charge	Identified Spectra	Annotated charge 1 b-ions		Annotated charge 1 y-ions	
			Before	After	Before	After
CID	2	562	4525	4661	4265	4418
	3	408	2325	3664	2440	3557
	4	210	933	1816	794	1922
	5	93	336	889	339	959
	6	43	177	483	87	452
	7	7	26	68	7	63
HCD	2	362	2761	2696	3262	3236
	3	534	3790	4011	4275	4736
	4	298	1894	2119	2453	3052
	5	117	814	993	1052	1311
	6	59	350	485	552	762
	7	16	95	129	151	223

Table S-1 - Validation of MS/MS Deconvolution Results: Annotated b and y ions were counted in identified CID and HCD spectra of precursor charge 2-7 from the 6-prot dataset both before and after MS/MS Deconvolution. Spectra were identified with SpectrumMill with FDR 1% before MS/MS Deconvution as described in *6-prot SpectrumMill Parameters*. The parameter β was then trained to maximize observed counts of charge 1 b and y ions in CID spectra ($\beta = 0.53$), where b and y ion counts were most improved by MS/MS Deconvolution.

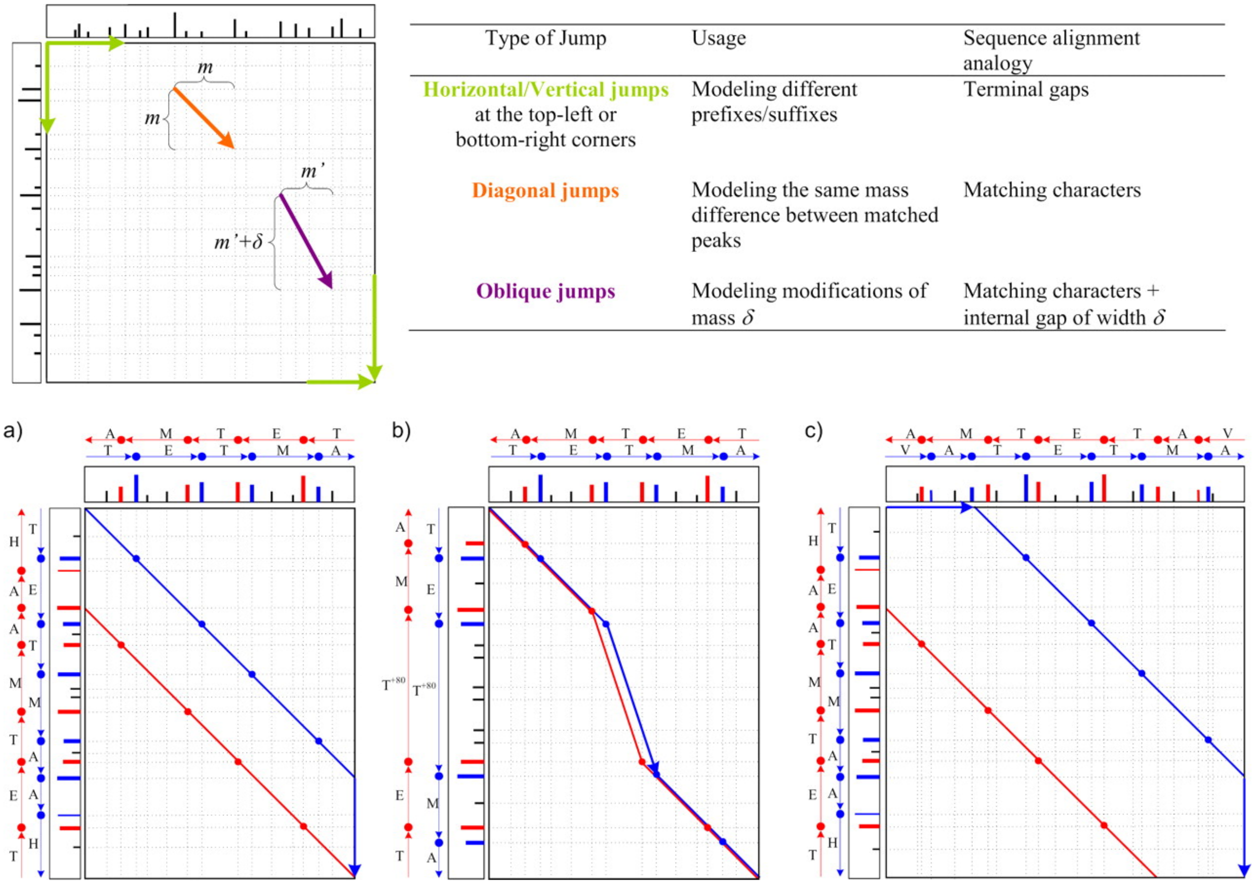
MS-GFDB Search Parameters

Deconvoluted spectra from each dataset were searched against reference protein sequences by MS-GFDB (6) v20110812 to validate de novo sequencing accuracy and coverage. The following parameters were set for all searches: carbamidomethylation (+57 Da) protecting group, 1 allowed ^{13}C, and non-specified protease. Furthermore, all searches were conducted allowing for the following modifications: oxidation of methionine, N-terminal pyroglutamate formation, deamidation of asparagine, and deamidation of aspartic acid. 6-prot and aBTLA Orbitrap CID spectra were searched with high accuracy LTQ instrument ID, CID fragmentation method, and 30 ppm parent mass tolerance. 6-prot HCD spectra were searched with high accuracy LTQ instrument, HCD fragmentation method, and 30 ppm parent mass tolerance. aBTLA Ion-trap spectra were searched with Ion-Trap instrument ID, CID fragmentation method, and 1.5 Da parent mass tolerance.

To accurately compute FDR for spectrum IDs, a large decoy database was appended to each small target database. All 6-prot searches were done on a database consisting of the reference sequences for the six target proteins, sequences of common contaminate proteins, and the same Drosophila proteome described in *6-prot SpectrumMill Parameters*. For the aBTLA searches the same contaminate sequences and Drosophila proteome were added to a database consisting of the reference sequences for the light and heavy chain of the aBTLA antibody. MS-GFDB then reported IDs at 1% spectrum-level FDR against the aggregated databases. Identified spectra were then separated by precursor charge and again filtered by 1% spectrum-level FDR as described in *6-prot SpectrumMill Parameters*.

Shotgun Protein Sequencing

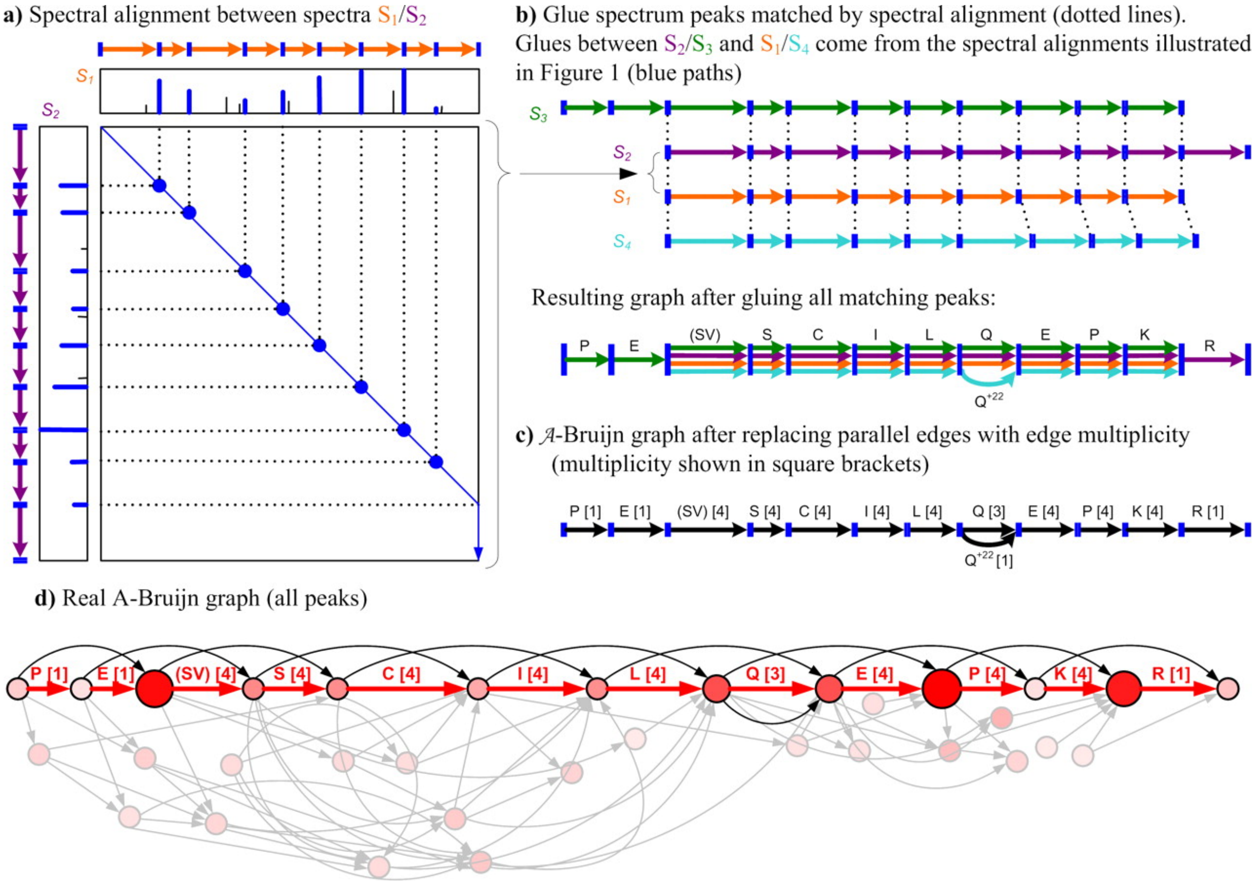
Figure S-1 (Reproduced from Figure 2 in (7))



Type of Jump	Usage	Sequence alignment analogy
Horizontal/Vertical jumps at the top-left or bottom-right corners	Modeling different prefixes/suffixes	Terminal gaps
Diagonal jumps	Modeling the same mass difference between matched peaks	Matching characters
Oblique jumps	Modeling modifications of mass δ	Matching characters + internal gap of width δ

(Caption for Figure S-1) Pairwise spectral alignments are computed with a dynamic programming algorithm similar to the Smith-Waterman sequence alignment algorithm; the corresponding intuitive interpretations are given in the table. The alignment of two spectra is defined on the set of all matching peaks; each pair of matching peaks is represented as an intersection of vertical and horizontal dotted lines on the spectral matrix (top left). 18 peaks in the first spectrum and 17 peaks in the second spectrum result in 17 × 18 matching peaks in the spectral matrix. Matching peaks may be connected by three types of jumps: horizontal/vertical (because MS/MS spectra commonly lack peaks in the low/high mass regions, we also accept horizontal/vertical jumps to locations where no peaks are matched), diagonal, and oblique jumps. A spectral alignment is defined as a sequence of jumps from the top left corner to the bottom right corner. We consider spectral alignments with any number of diagonal jumps but a limited number of other jumps and distinguish between three types of spectral alignments: **a)** prefix/suffix alignments use a single horizontal/vertical jump (either at the top left or bottom right); **b)** modified/unmodified alignments use a single oblique jump; and **c)** partial overlap alignments use one horizontal/vertical jump at the top left corner and another at the bottom right corner. The optimal alignment of two spectra is an alignment with the longest sequence of valid jumps on the spectral matrix (the implemented scoring function is described in the main text). The alignment of b ions is shown in blue, and that of y ions is shown in red.

Figure S-2 (Reproduced from Figure 3 in (7)**)**



a) Spectral alignment between spectra S_1/S_2

b) Glue spectrum peaks matched by spectral alignment (dotted lines). Glues between S_2/S_3 and S_1/S_4 come from the spectral alignments illustrated in Figure 1 (blue paths)

Resulting graph after gluing all matching peaks:

c) $\mathcal{A}$-Bruijn graph after replacing parallel edges with edge multiplicity (multiplicity shown in square brackets)

d) Real A-Bruijn graph (all peaks)

(Caption for Figure S-2) Construction of an A-Bruijn graph from aligned PRM spectra. Star spectra of peptides SVSCILQEPK (S1), SVSCILQEPKR (S2), PESVSCILQEPK (S3), and SVSCILQ+22EPK (S4) are "glued" together into an A-Bruijn graph using gluing instructions provided by pairwise spectral alignments shown in Figure 4. **a)** the spectral alignment of spectra S1 and S2 shown in Figure 4a reveals matching peaks in these spectra (only the blue path is shown). The peaks corresponding to b ions are shown in blue; other peaks are shown in black. Simplified spectrum graphs are shown next to each spectrum as paths through b ions. **b)** matching peaks in spectral alignments shown in Figure 4, a, b, and c, generate pairwise gluing instructions between every pair of aligned spectra. Thus, dotted lines are used to represent both matching peaks in a and gluing instructions in b. **c)** parallel edges are replaced by a single edge with weight proportional to its multiplicity. In reality, edge weights are determined from peak intensities. **d)** real A-Bruijn graph using all peaks in the aligned spectra. Vertex scores are represented as vertex size and color intensity; edges to noise peaks are shown in grey. The path found by shotgun protein sequencing is shown in red with edge labels for the identified amino acids (numbers in square brackets indicate edge multiplicity).

Meta-Assembly

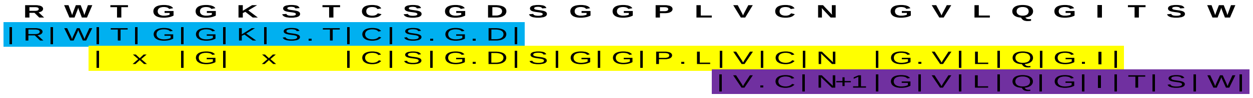
Input: Set of contig PRM spectra (S) and their alignments ($A\langle S_i, S_j\rangle$) where all alignments have score$\geq \tau$.
Output: Set of meta-contigs such that contig PRM spectra are consistent and coherent within each meta-contig.

```
1) Create an overlap graph $G = (V, E)$:
   a) For each contig PRM spectrum $S_i$ create a meta-contig $M_i$ containing only $S_i$ and add $M_i$ to $V$. Also create an empty set of alignments $Q_i$ for each $M_i$ that SPS will use when determining the meta-contig PRM spectrum of $M_i$
   b) For each alignment shift $A\langle S_i, S_j\rangle$, add the undirected alignment edge $e(M_i, M_j)$ to $E$ labeled with shift $A\langle M_i, M_j\rangle$, score$(e) \leftarrow$ score$(A)$, and reverse state R$[e] \leftarrow$ R$[A]$. $A\langle M_i, M_j\rangle$ is the shift of the meta-contig PRM spectrum of $M_j$ wrt the meta-contig PRM spectrum of $M_i$
Recruit
2) Recruit the highest scoring edge $e^*(M_i, M_j)$ with shift $A^*\langle M_i, M_j\rangle$. If score$(e^*) < \tau$ then halt computation and return all meta-contigs
Reverse
3) If R$[e^*] = true$ then reverse $M_j$ $(M_j = M_j^R)$:
   a) $S \leftarrow S^R \;\; \forall\, S \in M_j$
   b) $A \leftarrow A^R = \mathrm{PM}[S_a] - A - \mathrm{PM}[S_b] \;\; \forall\, A\langle S_a, S_b\rangle \in M_j$
   c) $A \leftarrow A^R \;\; \forall\, A\langle S_a, S_b\rangle \in Q_j$
   d) For each $e(M_j, M_k) \in E$ with shift $A\langle M_j, M_k\rangle$ st $k \neq i$, $A \leftarrow A^R$ and R$[A] \leftarrow not$ R$[A]$
Re-sequence
4) Create merged meta-contig $M_i^* \leftarrow (M_i \cup M_j)$, add it to $V$, and determine its meta-contig PRM spectrum:
   a) Add the inner edge $A\langle S_i, S_y\rangle \leftarrow A^* + A\langle S_j, S_y\rangle$ to $M_i^*$ for each $S_y \in M_j$ where $S_i$ was the first SPS contig PRM spectrum in $M_i$
   b) Create $Q_i^* \leftarrow (Q_i \cup Q_j)$. Over all unique pairs of contig PRM spectra $S_x \in M_i$ and $S_y \in M_j$, find the shift $A''\langle S_x, S_y\rangle = A\langle S_i, S_y\rangle - A\langle S_i, S_x\rangle$ with maximum MP($A''$) and add the alignment of $A''\langle S_x, S_y\rangle$ to $Q_i^*$
   c) Using SPS (8), sequence the meta-contig PRM spectrum of $M_i^*$ as the SPS consensus sequence of all contig PRM spectra in $M_i^*$ assembled by all alignments in $Q_i^*$
Re-score
5) Transfer, update, and re-score alignments to $M_i^*$:
   a) For each $M_k$ where $e_1(M_i, M_k) \in E$ and $e_2(M_j, M_k) \notin E$, add $e_1^*(M_i^*, M_k)$ to E with the same labels as $e_1$ (its shift is $A_1^*\langle M_i^*, M_k\rangle = A_1\langle M_i, M_k\rangle$)
   b) For each $M_k$ where $e_1(M_i, M_k) \notin E$ and $e_2(M_j, M_k) \in E$, add $e_2^*(M_i^*, M_k)$ to E with the same labels as $e_2$ except its shift $A_2^*\langle M_i^*, M_k\rangle = A^* + A_2$.
   c) For each $M_k$ where $e_1(M_i, M_k) \in E$ and $e_2(M_j, M_k) \in E$, consider edges $e_1^*$ and $e_2^*$ with shifts $A_1^*$ and $A_2^*$ as done in steps 5a and 5b, respectively. If score$(A_1^*) >$ score$(A_2^*)$, then add $e_1^*$ to $E$ as done in step 5a. Otherwise, add $e_2^*$ to $E$ as done in step 5b.
   d) For each $M_k$ connected to $M_i^*$ through an edge $e(M_i^*, M_k)$ labeled with shift $A\langle M_i^*, M_k\rangle$, re-label score$(e) \leftarrow$ score$(A)$ as the score of the alignment between the meta-contig PRM spectra of $M_i^*$ and $M_k$. If R$[e] = true$ then temporarily reverse the meta-contig PRM spectrum of $M_k$ when computing the score.
   e) Remove $M_i$, $M_j$, and their edges from $G$
6) Iterate to step 2
```

Supplementary Results

Running Time – Meta-SPS was implemented in C++ and compiled on g++ version 4.4.3 (Ubuntu Linux x64) with the –O3 optimization flag. The system included a 3.20 GHz Intel Core i7 CPU (model 960) and 12 GB of available RAM. For the alignment and assembly of 6-prot contigs, Spectral Alignment ran in 8 minutes, 23 seconds when multiplexed on 7 parallel threads (using the pthread library). Running on one thread, Meta-Assembly ran in 4.5 seconds. For both steps, Meta-SPS used no more than 200 MB of memory.

Figure S-3



```
 R W T G G K S T C S G D S G G P L V C N   G V L Q G I T S W
|R|W|T|G|G|K|S.T|C|S.G.D|
    |  x  |G|  x  |C|S|G.D|S|G|G|P.L|V|C|N  |G.V|L|Q|G.I|
                                     |V.C|N+1|G|V|L|Q|G|I|T|S|W|
```

Figure S-3 – Overlap of un-joined meta contig sequences in kallikrein-related peptidase: This details the overlap of the light blue, yellow, and purple meta-contigs displayed in Figure 4a. Amino acid letters correspond to mapped sequence calls and gaps. Each "|" corresponds to an expected PRM mass that is present in the meta-contig PRM spectrum. Each "." corresponds to an expected PRM mass that is absent in the meta-contig PRM spectrum (although the surrounding gap may still be correct). Each "x" corresponds to an expected PRM mass that is present in the meta-contig PRM spectrum at the incorrect mass (leading to incorrect sequence calls). The "N+1" corresponds to a deamidation of Asn that was correctly called in the purple meta-contig sequence. Because of 3 missing PRMs in the blue meta-contig, 2 incorrect PRMs in the yellow meta-contig, and 1 missing PRM in the yellow meta-contig, the yellow and blue meta-contigs only shared 3 matching peaks and were not merged (6 were required). Because of 2 missing PRMs in the yellow meta-contig, 1 missing PRM in the purple meta-contig, and the N+1 modification in the purple meta-contig, the yellow and purple meta-contigs only shared 5 matching peaks and were not merged.

Table S-2

		6-Protein Mixture						aBLTA	
	Protein	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
	Protein Length (AA)	167	261	548	154	100	353	219	443
	Spectrum Coverage (%)	94.6	90.8	97.8	99.4	60.0	67.7	97.7	100
Coverage	Mapped Meta-contigs	14	18	49	12	9	30	11	28
	Sequencing Coverage (%)	83.8	84.7	84.3	83.8	56	52.4	87.7	93.7
	Coverage Redundancy	1.5	1.5	1.5	1.3	1.6	1.5	1.3	1.3
	Spectra Per Meta-contig	48	69	39	20	41	22	96	58
	Peptides Per Meta-contig	15	18	13	7	11	8	34	31
	Average Seq. Length (AA)	14.6	18.3	14.3	13.7	9.8	9.5	22.3	18.6
	Longest Sequence (AA)	45	91	47	25	17	21	106	60
Accuracy	Identified Meta-contigs	10	12	29	8	5	14	7	23
	Correct Sequence Calls (%)	85.9	84.9	84.9	92.4	87.0	97.4	98.4	95.8
	Un-annotated Seq. Calls (%)	14.9	2.1	4.0	0.8	11.5	1.3	3.2	9.3

(Caption for Table S-2) Meta-contig sequencing coverage, length, and accuracy including un-merged SPS contigs: Protein identifiers are: P_1 - leptin precursor, P_2 – kallikrein-related peptidase, P_3 – GroEL, P_4 – myoglobin, P_5 – aprotinin, P_6 – peroxidase, P_7 – aBTLA light chain, and P_8 – aBTLA heavy chain. *Protein Length* is the length of each reference protein in amino acid residues. *Spectrum Coverage* is the percent of each spectrum covered by peptides identified MS-GFDB with 1% FDR. *Coverage* is taken over all mapped contigs and *Accuracy* is taken over all identified meta-contigs. Mapped meta-contigs must be aligned to a reference protein as described in the text while identified meta-contigs must assemble at least one identified spectrum whose peptide sequence is a substring of a reference protein. *Sequencing Coverage* is the percent of amino acids in each protein covered by at least one mapped meta-contig sequence. *Coverage Redundancy* is the average number of mapped meta-contig sequences covering each amino acid residue that is covered by at least one meta-contig sequence. *Spectra Per Meta-contig* is the average number of spectra assembled by each mapped meta-contig while *Peptides Per Meta-contig* is the average number of peptides (spectra with distinct parent masses) assembled by each mapped meta-contig. *Average Seq. Length* is the average number of amino acid residues covered by each mapped meta-contig and *Longest Sequence* is the maximum number of amino acid residues covered by a mapped meta-contig. *Correct Sequence Calls* is the percentage of annotated sequence calls that were correct in identified meta-contigs. *Un-annotated Seq. Calls* is the percentage of sequence calls that were un-annotated in identified meta-contigs.

Table S-3

		6-Protein Mixture						aBLTA	
	Protein	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
	Protein Length (AA)	167	261	548	154	100	353	219	443
	Spectrum Coverage (%)	94.6	90.8	97.8	99.4	60.0	67.7	97.7	100
Coverage	Mapped Contigs	60	104	141	30	20	61	50	111
	Sequencing Coverage (%)	90.4	88.9	85.6	84.4	54.0	56.1	90	95.9
	Coverage Redundancy	4.2	4.7	3.7	2.5	3.4	2.9	2.9	3.1
	Spectra Per Contig	13	12	13	7	18	10	20	12.4
	Peptides Per Contig	7	6	6	4	7	5	12	9
	Average Seq. Length (AA)	10.7	10.6	12.4	11	9.2	9.4	11.4	12.0
	Longest Sequence (AA)	24	26	32	16	18	19	36	33
Accuracy	Identified Contigs	42	69	113	24	13	44	33	19
	Correct Sequence Calls (%)	86.8	85.6	87.2	95.1	88.7	93.7	98.7	95.6
	Un-annotated Seq. Calls (%)	7.9	2.3	3.7	0.4	5.3	1.7	4	3.8

(Caption for Table S-3) SPS sequencing coverage, length, and accuracy: Protein identifiers are: P_1 - leptin precursor, P_2 – kallikrein-related peptidase, P_3 – GroEL, P_4 – myoglobin, P_5 – aprotinin, P_6 – peroxidase, P_7 – aBTLA light chain, and P_8 – aBTLA heavy chain. *Protein Length* is the length of each reference protein in amino acid residues. *Spectrum Coverage* is the percent of each spectrum covered by peptides identified MS-GFDB with 1% FDR. *Coverage* is taken over all mapped contigs and *Accuracy* is taken over all identified contigs. Mapped contigs must be aligned to a reference protein as described in the text while identified contigs must assemble at least one identified spectrum whose peptide sequence is a substring of a reference protein. *Sequencing Coverage* is the percent of amino acids in each protein covered by at least one mapped contig sequence. *Coverage Redundancy* is the average number of mapped contig sequences covering each amino acid residue that is covered by at least one contig sequence. *Spectra Per Contig* is the average number of spectra assembled by each mapped contig while *Peptides Per Contig* is the average number of peptides (spectra with distinct parent masses) assembled by each mapped contig. *Average Seq. Length* is the average number of amino acid residues covered by each mapped contig and *Longest Sequence* is the maximum number of amino acid residues covered by a mapped contig. *Correct Sequence Calls* is the percentage of annotated sequence calls that were correct in identified contigs. *Un-annotated Seq. Calls* is the percentage of sequence calls that were un-annotated in identified contigs.

Table S-4

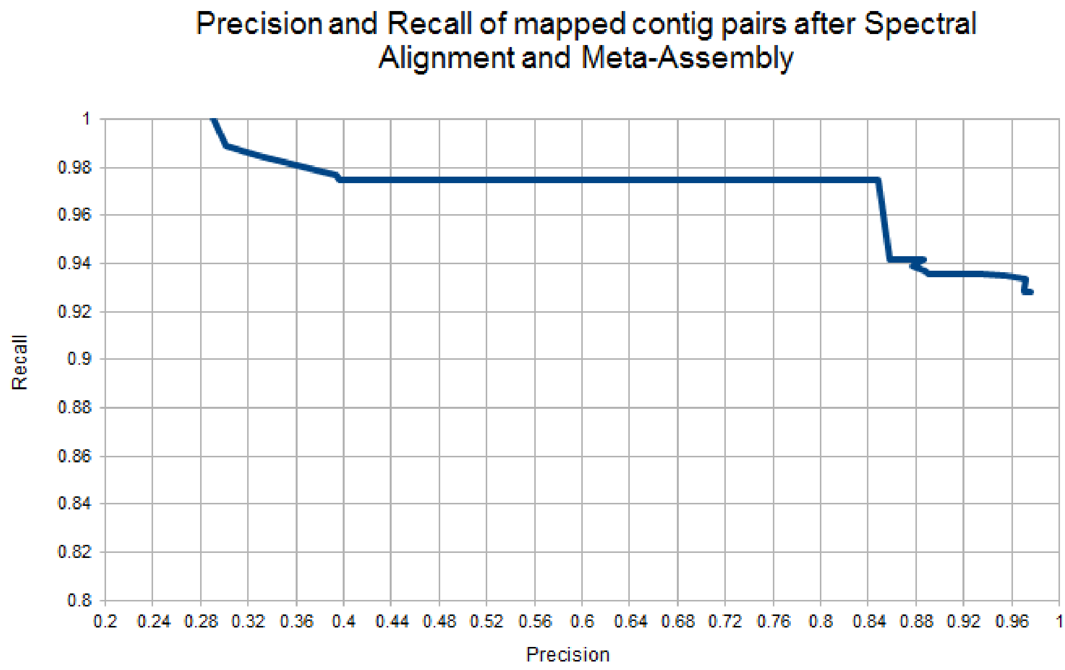
		aBLTA FT CID spectra only		Mixed aBTLA/6prot CID FT spectra	
	Protein	P_7	P_8	P_7	P_8
	Protein Length (AA)	219	443	219	443
	Spectrum Coverage (%)	97.7	98	97.7	98
Coverage	Mapped Contigs	10	16	7	14
	Sequencing Coverage (%)	61.2	58.5	58	58.9
	Coverage Redundancy	1	1.1	1	1
	Spectra Per Contig	26.6	31.5	32.6	38.7
	Peptides Per Contig	18.8	22.7	20.3	26.9
	Average Seq. Length (AA)	13.7	18.4	18.3	19.3
	Longest Sequence (AA)	25	37	24	35
Accuracy	Identified Contigs	9	17	8	15
	Correct Sequence Calls (%)	100	97.7	98.3	98
	Un-annotated Seq. Calls (%)	3	5.5	6.2	2.4

(Caption for Table S-4) Meta-SPS sequencing coverage, length, and accuracy for the aBTLA/6-prot mixture: Protein identifiers are: P_7 – aBTLA light chain and P_8 – aBTLA heavy chain. *Protein Length* is the length of each reference protein in amino acid residues. *Spectrum Coverage* is the percent of each spectrum covered by peptides identified MS-GFDB with 1% FDR. *Coverage* is taken over all mapped contigs and *Accuracy* is taken over all identified contigs. Mapped contigs must be aligned to a reference protein as described in the text while identified contigs must assemble at least one identified spectrum whose peptide sequence is a substring of a reference protein. *Sequencing Coverage* is the percent of amino acids in each protein covered by at least one mapped contig sequence. *Coverage Redundancy* is the average number of mapped contig sequences covering each amino acid residue that is covered by at least one contig sequence. *Spectra Per Contig* is the average number of spectra assembled by each mapped contig while *Peptides Per Contig* is the average number of peptides (spectra with distinct parent masses) assembled by each mapped contig. *Average Seq. Length* is the average number of amino acid residues covered by each mapped contig and *Longest Sequence* is the maximum number of amino acid residues covered by a mapped contig. *Correct Sequence Calls* is the percentage of annotated sequence calls that were correct in identified contigs. *Un-annotated Seq. Calls* is the percentage of sequence calls that were un-annotated in identified contigs. The left two columns detail sequencing statistics for Meta-SPS using only the aBTLA high resolution CID spectra. The right two columns detail the same statistics for meta-contigs using aBTLA high resolution CID spectra combined with the 6-prot high resolution CID spectra.

Using only the aBTLA high resolution CID spectra, the target set of proteins was small enough such that combining meta-contigs with un-merged SPS contigs yielded the best overall combination of sequencing accuracy, length, and coverage. To optimize results for the aBTLA/6-prot mixture, SPS sequencing parameters had to be adjusted to better control false positive alignments. Since the resulting contigs were smaller and less accurate, only meta-contigs assembling at least two SPS contigs or more were allowed. Thus the increase in average sequence length from aBTLA meta-contigs/contigs (left two columns) to aBTLA/6-prot meta-contigs (right two columns) is explained by the incorporation of shorter SPS contigs into the aBTLA sequencing results.

Explanation of incorrect meta-contig sequence calls at positions k = 20, 21, and 22: In this case there was little overlapping coverage of MS/MS spectra (and contig PRM spectra) in the meta-contig at those positions. This type of coverage is typically observed towards the ends of meta-contigs. But here the three incorrect sequence calls in the meta-contig were inherited from the middle of a 26 AA long SPS contig (very long for a SPS contig) that spanned a region with only one more overlapping contig PRM spectrum.

Figure S-4



(Caption for Figure S-4) Meta-assembly Precison/Recall: Precision (ratio of the number of mapped contig PRM spectral pairs in the same meta-contig whose observed shift matches the theoretical within parent mass tolerance or a PTM mass over the number of mapped contig PRM spectral pairs in the same meta-contig) and recall (ratio of the number of all mapped contig PRM spectral pairs with at least 6 matching peaks that were merged into a meta-contig over the number of all mapped contig PRM spectral pairs with at least 6 matching peaks) was computed for separate runs of Meta-SPS at various values of τ. In both data sets, τ was trained to achieve 97% precision (curve for 6-prot is shown here).

Figure S-5

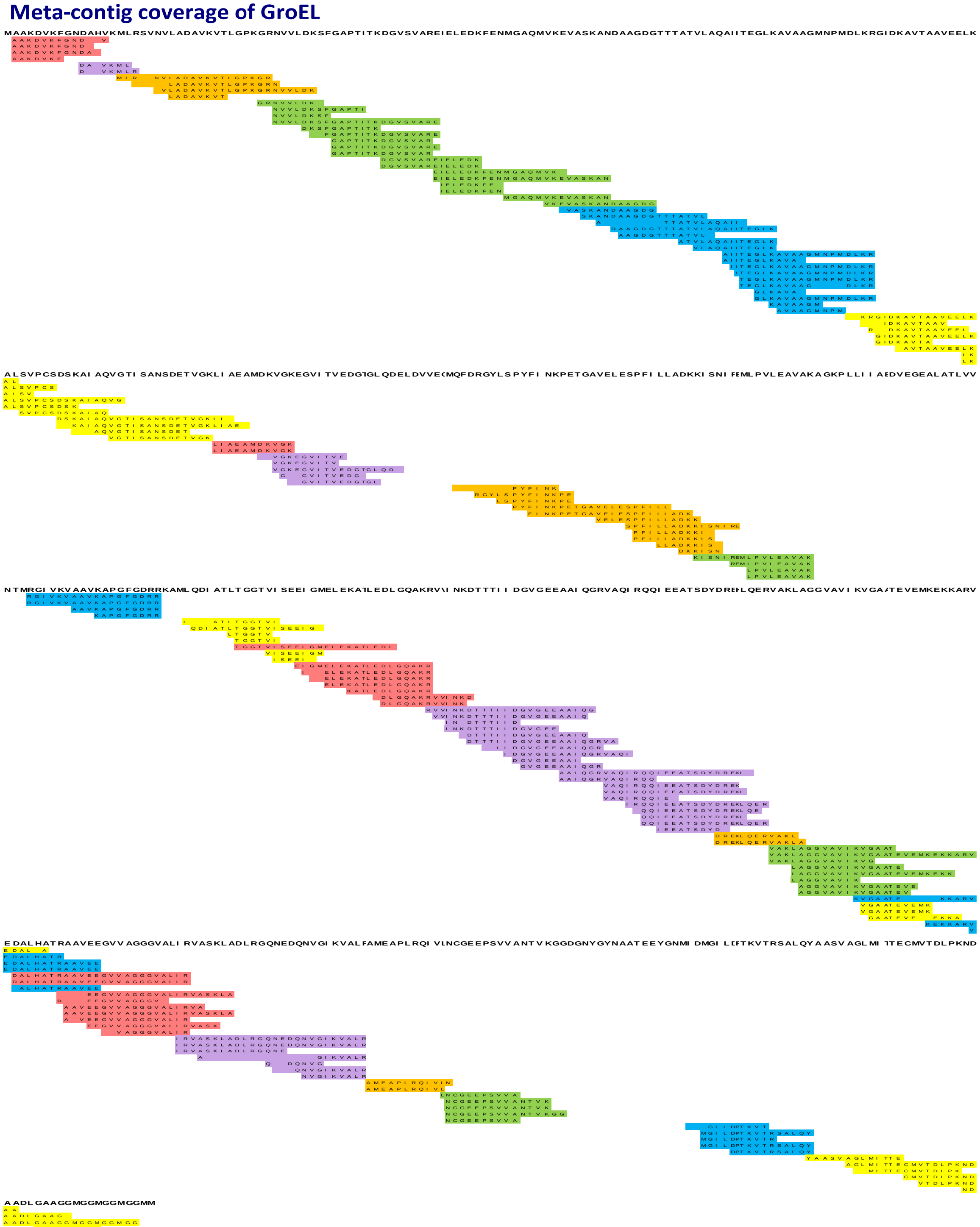

Meta-contig coverage of GroEL
MAAKDVKFGNDAHVKMLRSVNVLADAVKVTLGPKGRNVVLDKSFGAPTITKDGVSVAREIELEDKFENMGAQMVKEVASKANDAAGDGTTTATVLAQAIITEGLKAVAAGMNPMDLKRGIDKAVTAAVEELK
ALSVPCSDSKAIAQVGTISANSDETVGKLIAEAMDKVGKEGVITVEDGTGLQDELDVVEGMQFDRGYLSPYFINKPETGAVELESPFILLADKKISNIREMLPVLEAVAKAGKPLLIIAEDVEGEALATLVV
NTMRGIVKVAAVKAPGFGDRRKAMLQDIATLTGGTVISEEIGMELEKATLEDLGQAKRVVINKDTTTIIDGVGEEAAIQGRVAQIRQQIEEATSDYDREKLQERVAKLAGGVAVIKVGAATEVEMKEKKARV
EDALHATRAAVEEGVVAGGGVALIRVASKLADLRGQNEDQNVGIKVALRAMEAPLRQIVLNCGEEPSVVANTVKGGDGNYGYNAATEEYGNMIDMGILDPTKVTRSALQYAASVAGLMITTECMVTDLPKND
AADLGAAGGMGGMGGMGGMM

Figure S-5 – Meta-contig coverage of GroEL from the 6-prot sample is displayed here (as in Figure 4). Contigs that overlap but were not merged are explained by circumstances illustrated in Figure 6.

Figure S-6

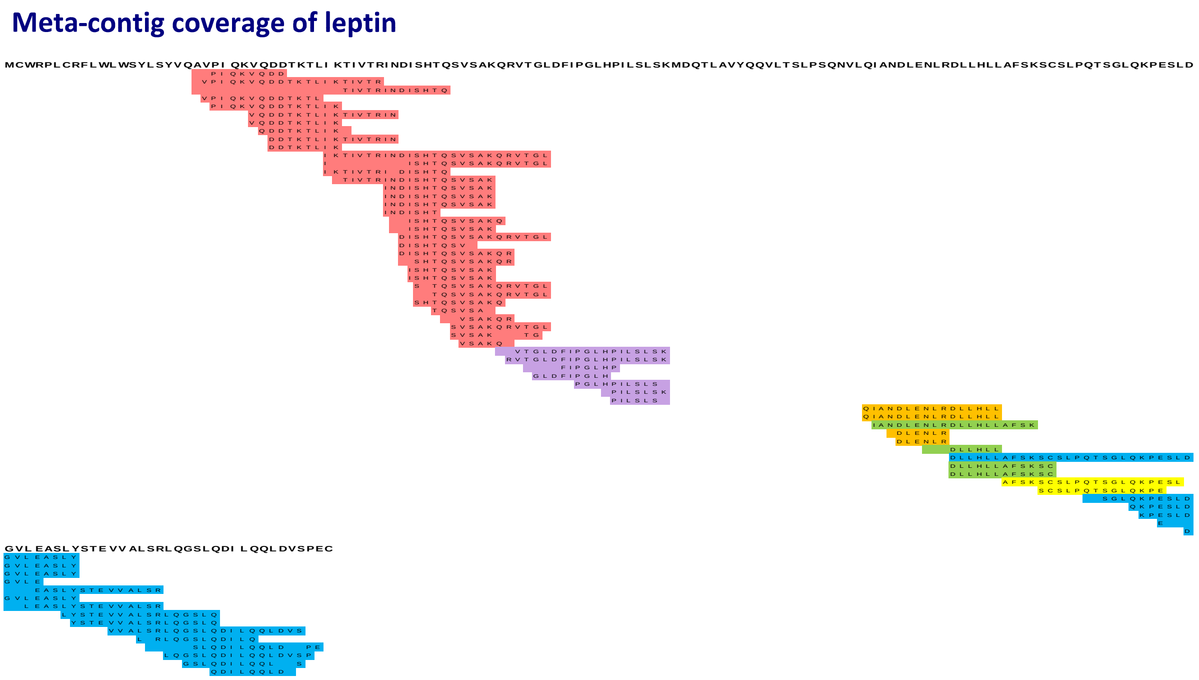

Meta-contig coverage of leptin
MCWRPLCRFLWLWSYLSYVQAVPI QKVQDDTKTLI KTIVTRINDISHTQSVSAKQRVTGLDFIPGLHPILSLSKMDQTLAVYQQVLTSLPSQNVLQIANDLENLRDLLHLLAFSKSCSLPQTSGLQKPESLD
GVLEASLYSTEVVALSRLQGSLQDI LQQLDVSPEC

Figure S-6 – Meta-contig coverage of leptin from the 6-prot sample is displayed here (as in Figure 4). Contigs that overlap but were not merged are explained by circumstances illustrated in Figure 6.

Figure S-7

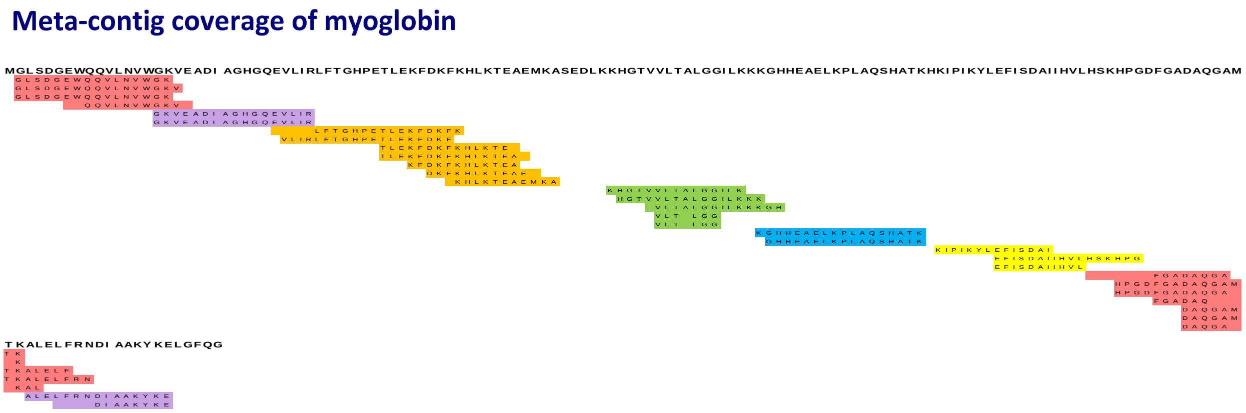

Meta-contig coverage of myoglobin
MGLSDGEWQQVLNVWGKVEADI AGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDLKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKHPGDFGADAQGAM
TKALELFRNDI AAKYKELGFQG

Figure S-7 – Meta-contig coverage of myoglobin from the 6-prot sample is displayed here (as in Figure 4). Contigs that overlap but were not merged are explained by circumstances illustrated in Figure 6.

Figure S-8

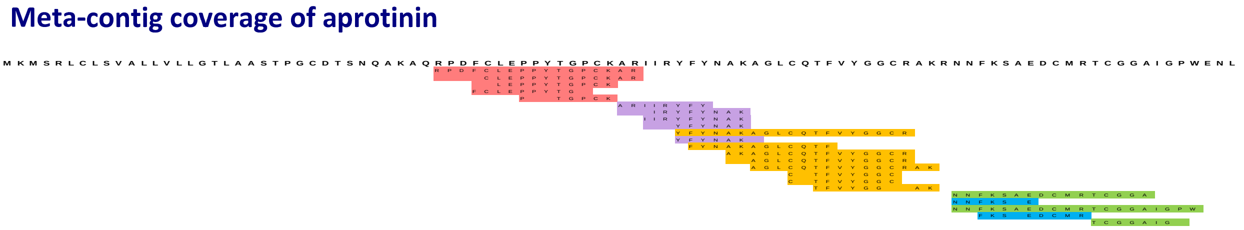

Meta-contig coverage of aprotinin
M K M S R L C L S V A L L V L L G T L A A S T P G C D T S N Q A K A Q R P D F C L E P P Y T G P C K A R I I R Y F Y N A K A G L C Q T F V Y G G C R A K R N N F K S A E D C M R T C G G A I G P W E N L

Figure S-8 – Meta-contig coverage of aprotinin from the 6-prot sample is displayed here (as in Figure 4). Contigs that overlap but were not merged are explained by circumstances illustrated in Figure 6.

Figure S-9

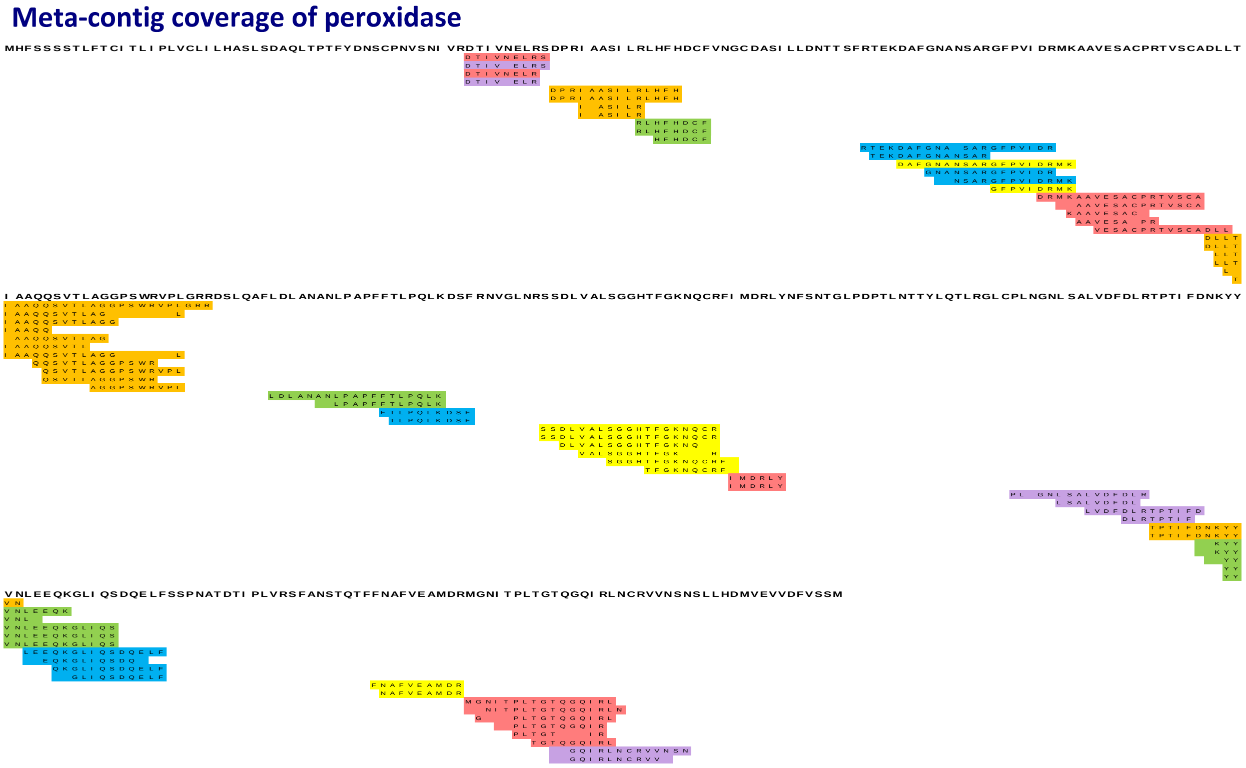

Meta-contig coverage of peroxidase
MHFSSSSTLFTCI TLI PLVCLI LHASLSDAQLTPTFYDNSCPNVSNI VRDTI VNELRSDPRI AASI LRLHFHDCFVNGCDASI LLDNTT SFRTEKDAFGNANSARGFPVI DRMKAAVESACPRTVSCADLLT
I AAQQSVTLAGGPSWRVPLGRRDSLQAFLDLANANLPAPFFTLPQLKDSFRNVGLNRSSDLVALSGGHTFGKNQCRFI MDRLYNFSNTGLPDPTLNTTYLQTLRGLCPLNGNL SALVDFDLRTPTI FDNKYY
VNLEEQKGLI QSDQELFSSPNATDTI PLVRSFANSTQTFFNAFVEAMDRMGNI TPLTGTQGQI RLNCRVVNSNSLLHDMVEVVDFVSSM

Figure S-9 – Meta-contig coverage of peroxidase from the 6-prot sample is displayed here (as in Figure 4). Contigs that overlap but were not merged are explained by circumstances illustrated in Figure 6.

Figure S-10

Meta-contig coverage of aBTLA heavy chain

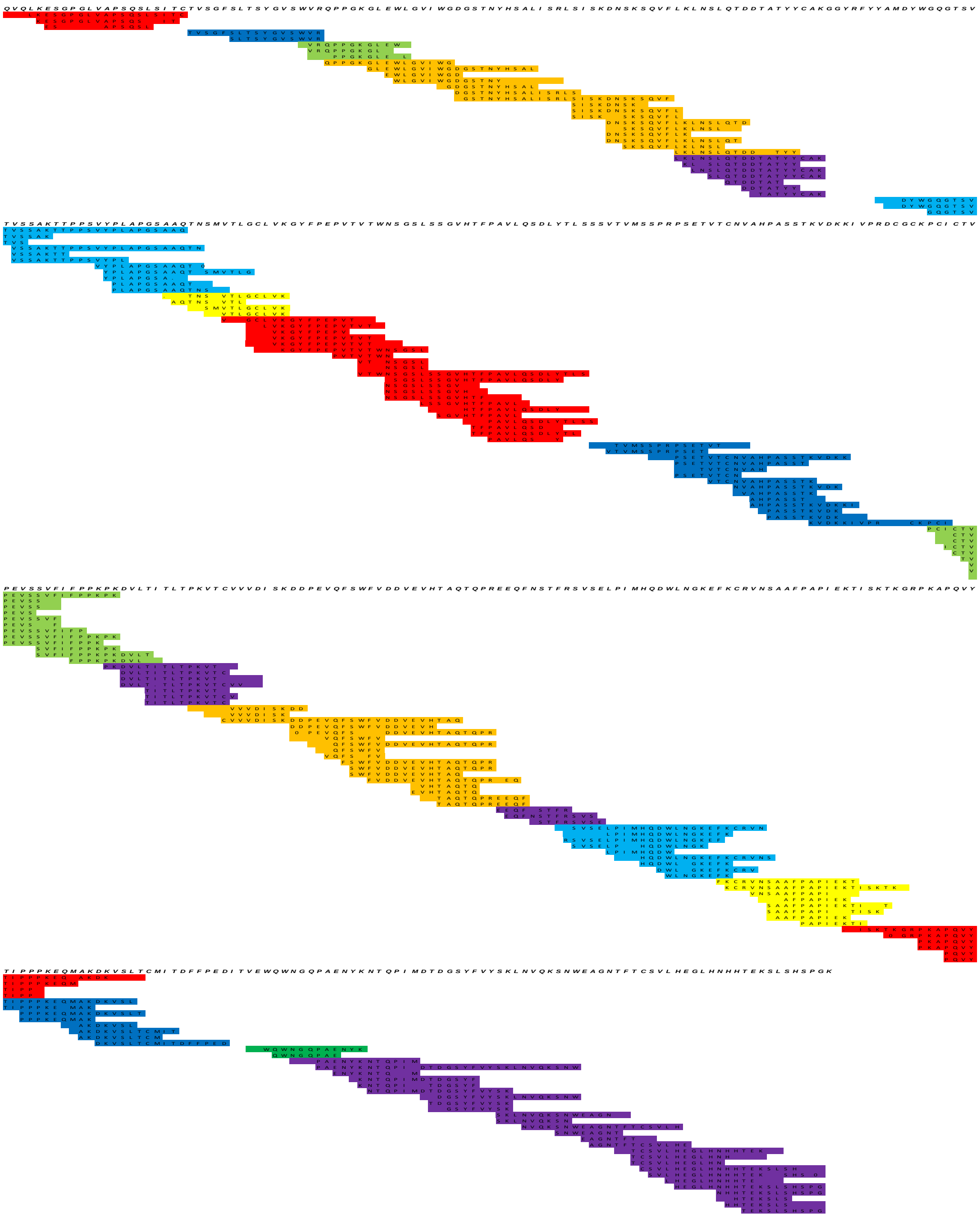

Q V Q L K E S G P G L V A P S Q S L S I T C T V S G F S L T S Y G V S W V R Q P P G K G L E W L G V I W G D G S T N Y H S A L I S R L S I S K D N S K S Q V F L K L N S L Q T D D T A T Y Y C A K G G Y R F Y Y A M D Y W G Q G T S V
T V S S A K T T P P S V Y P L A P G S A A Q T N S M V T L G C L V K G Y F P E P V T V T W N S G S L S S G V H T F P A V L Q S D L Y T L S S S V T V M S S P R P S E T V T C N V A H P A S S T K V D K K I V P R D C G C K P C I C T V
P E V S S V F I F P P K P K D V L T I T L T P K V T C V V V D I S K D D P E V Q F S W F V D D V E V H T A Q T Q P R E E Q F N S T F R S V S E L P I M H Q D W L N G K E F K C R V N S A A F P A P I E K T I S K T K G R P K A P Q V Y
T I P P P K E Q M A K D K V S L T C M I T D F F P E D I T V E W Q W N G Q P A E N Y K N T Q P I M D T D G S Y F V Y S K L N V Q K S N W E A G N T F T C S V L H E G L H N H H T E K S L S H S P G K

Figure S-10 – Meta-contig coverage of the aBTLA heavy-chain is displayed here (as in Figure 4). Contigs that overlap but were not merged are explained by circumstances illustrated in Figure 6.

Figure S-11 A)

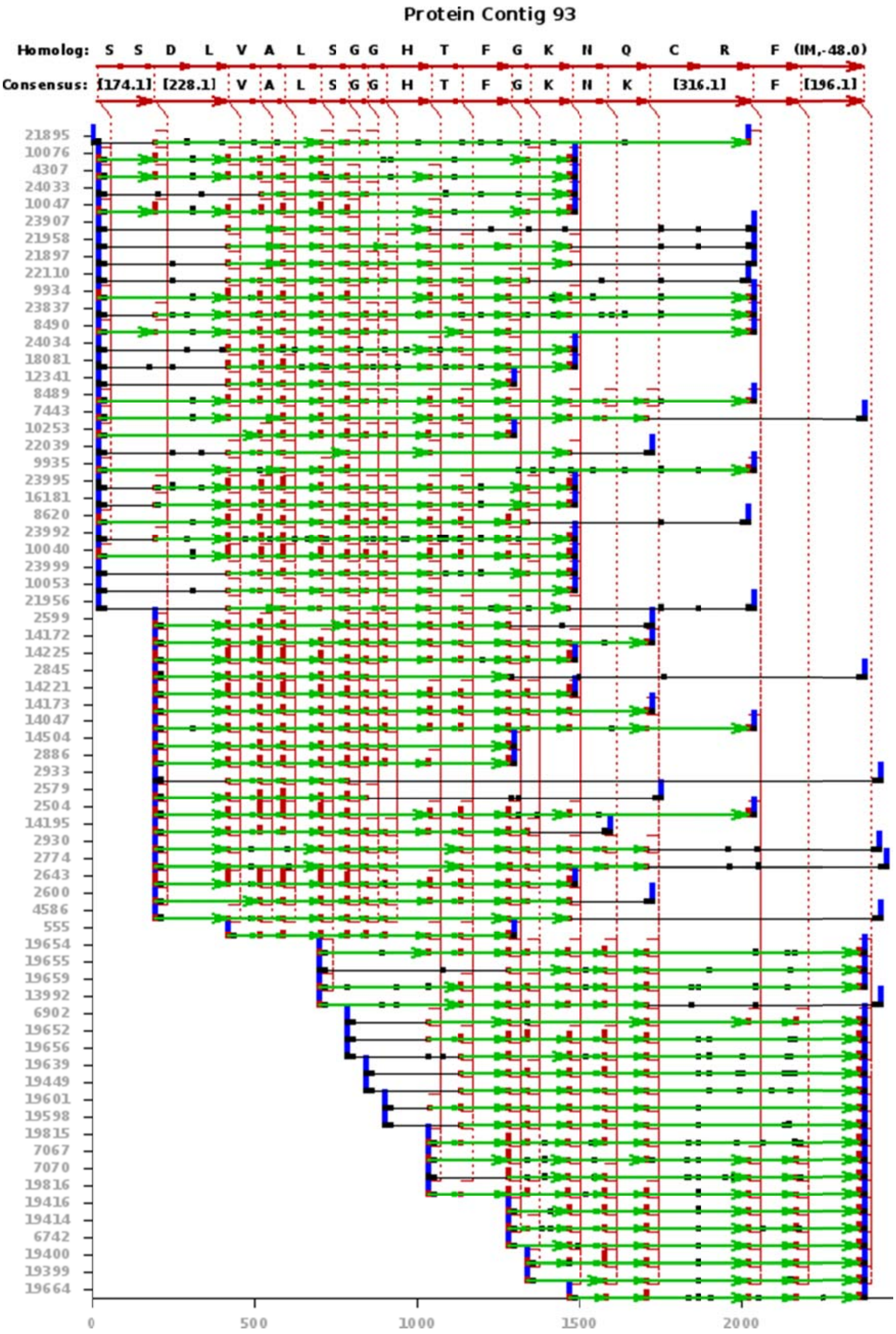
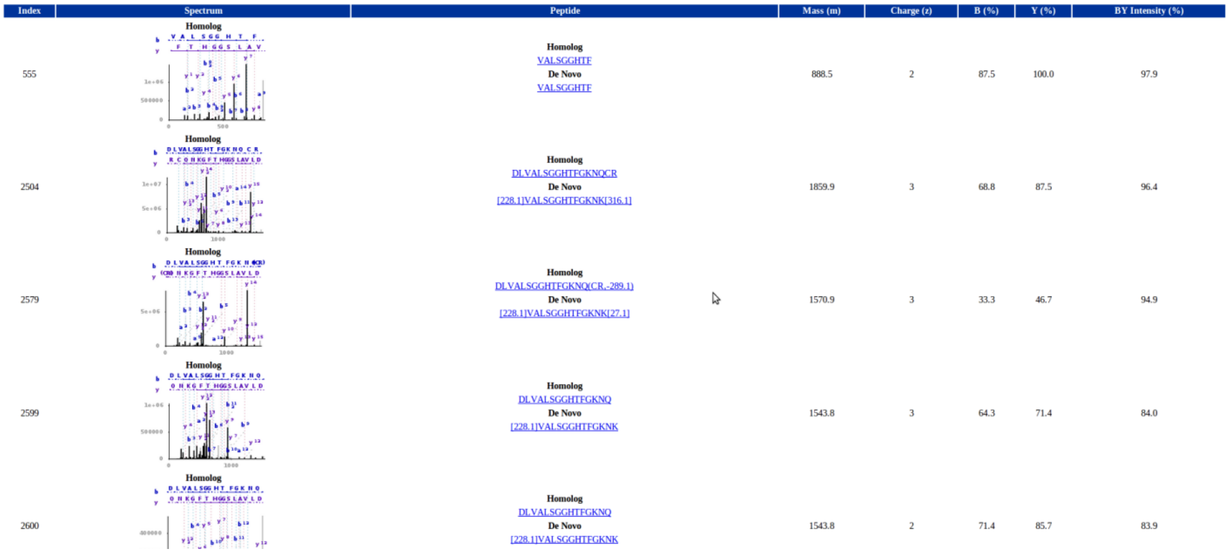
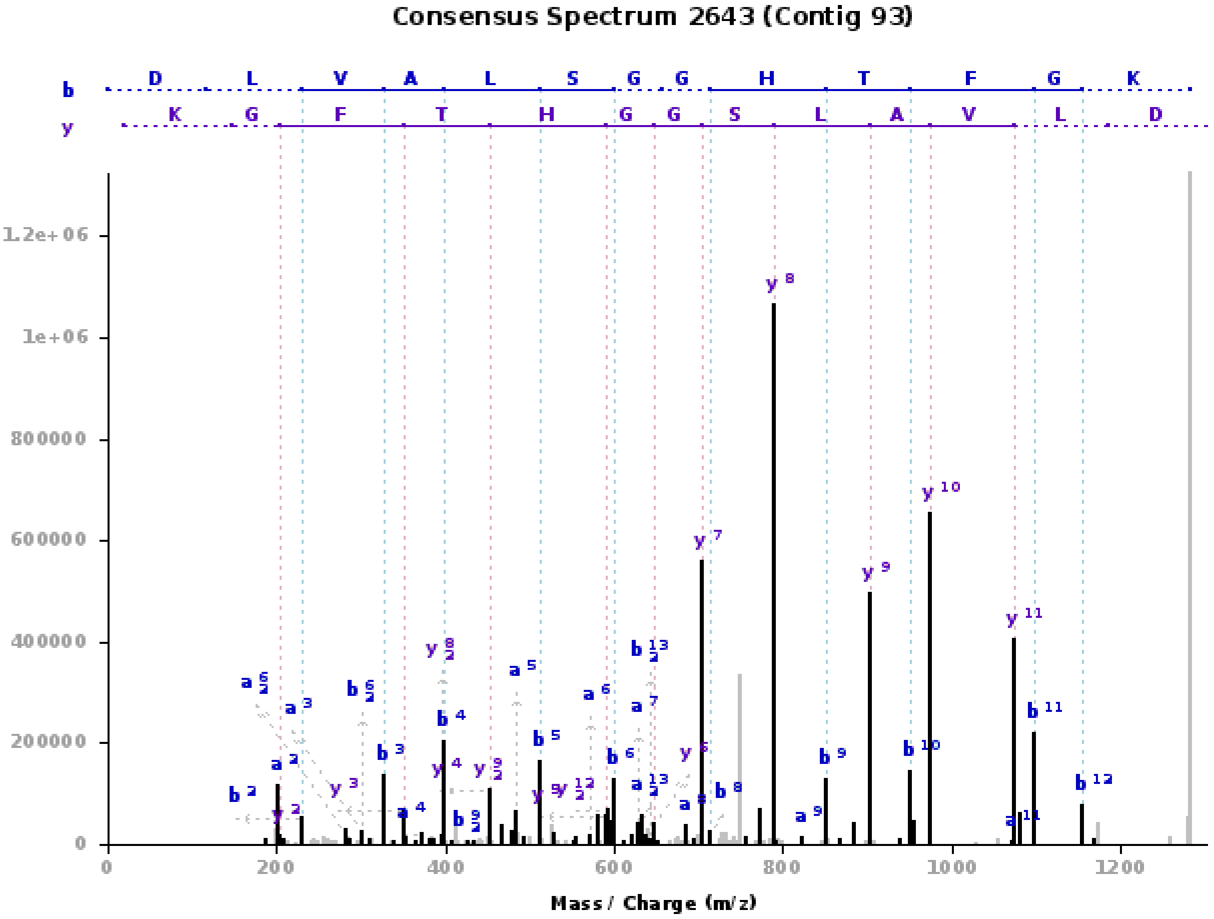

Protein Contig 93
Homolog: S S D L V A L S G G H T F G K N Q C R F (IM,-48.0)
Consensus: [174.1] [228.1] V A L S G G H T F G K N K [316.1] F [196.1]
21895
10076
4307
24033
10047
23907
21958
21897
22110
9934
23837
8490
24034
18081
12341
8489
7443
10253
22039
9935
23995
16181
8620
23992
10040
23999
10053
21956
2599
14172
14225
2845
14221
14173
14047
14504
2886
2933
2579
2504
14195
2930
2774
2643
2600
4586
555
19654
19655
19659
13992
6902
19652
19656
19639
19449
19601
19598
19815
7067
7070
19816
19416
19414
6742
19400
19399
19664
0
500
1000
1500
2000

Figure S-11 B)



Index	Spectrum	Peptide	Mass (m)	Charge (z)	B (%)	Y (%)	BY Intensity (%)
555	Homolog	**Homolog** VALSGGHTF **De Novo** VALSGGHTF	888.5	2	87.5	100.0	97.9
2504	Homolog	**Homolog** DLVALSGGHTFGKNQCR **De Novo** [228.1]VALSGGHTFGKNK[316.1]	1859.9	3	68.8	87.5	96.4
2579	Homolog	**Homolog** DLVALSGGHTFGKNQ(CR,-289.1) **De Novo** [228.1]VALSGGHTFGKNK[27.1]	1570.9	3	33.3	46.7	94.9
2599	Homolog	**Homolog** DLVALSGGHTFGKNQ **De Novo** [228.1]VALSGGHTFGKNK	1543.8	3	64.3	71.4	84.0
2600	Homolog	**Homolog** DLVALSGGHTFGKNQ **De Novo** [228.1]VALSGGHTFGKNK	1543.8	2	71.4	85.7	83.9

Figure S-11 C)



(Caption for Figure S-11) De novo sequencing reports: Selected screenshots from de novo sequencing reports uploaded to Tranche (see Results section for link). Reports for 6-prot meta-contigs are also directly available at http://proteomics.ucsd.edu/Software/MetaSPS/6-prot_meta-contigs/index.html, while this particular example can be viewed at http://proteomics.ucsd.edu/Software/MetaSPS/6-prot_meta-contigs/contig.93.html. All contigs and meta-contigs are listed as thumbnails similar to (A). After clicking on a contig's thumbnail, its full-size image is displayed along with a list of its assembled MS/MS spectra (as in (B)) where each spectrum is annotated with the de novo as well as homolog sequence (best match to a database sequence after completing de novo sequencing – only used for performance assessment purposes). After clicking on an annotated spectrum, its full-size image is displayed (as in (C)). **(A)** At the top is the database-mapped homolog sequence aligned to the de novo sequence. Below the homolog is the consensus, or de novo, sequence that Meta-SPS extracted from the assembled spectra. Below the de novo sequence are all the scored PRM spectra that were assembled into the contig. Horizontal green arrows denote amino acid jumps between PRMs that contributed to the consensus de novo sequence. Vertical red dotted lines detail which PRMs were grouped together by spectrum/spectrum and contig/contig alignment stages in SPS and Meta-SPS, respectively. Spectrum indices at the far left match those in (B), so one may view the MS/MS spectrum for each assembled PRM spectrum. Note that the homolog match labeled "(IM,-48)" at the far right indicates a verified PTM of -48 Da (homoserine lactone formation as a result of CNBr digestion). Though not the focus of the Meta-SPS algorithm presented here, we note that this PTM was confirmed by MS-GFDB at 1% spectrum-level FDR. **(B)** List of annotated MS/MS spectra for each assembled PRM spectrum in (A). At the far left are spectrum indices matching those in (A). The next column going to the right shows a thumbnail of each MS/MS spectrum annotated with the database-mapped homologous sequence from (A). Clicking on a thumbnail opens another page with the same full-size image (as in (C)). The next column contains links to the spectrum annotated by the de novo and homolog sequences. Remaining columns contain the precursor mass of the spectrum, the precursor charge of the spectrum, the percent of breaks observed by b/y ions, and the percent of intensity in b or y ions. **(C)** A full-size image of an assembled MS/MS spectrum annotated by the consensus de novo sequence.

From these reports, one can observe how de novo sequences are extracted from unidentified MS/MS spectra via the workflow described in Figure 1a. In this example, CID spectrum 2643 (Figure S-11C) entered the Meta-SPS pipeline as a cluster of unidentified MS/MS scans (clusters were generated by SpectrumMill for this data set). SPS first invoked PepNovo to convert spectrum 2643 into a PRM spectrum. The PRM spectrum was then aligned to every other spectrum in the dataset and grouped into a component of aligned spectra from overlapping peptides. SPS extracted from that component a contig PRM spectrum, which was further aligned to other overlapping contig PRM spectra by Meta-SPS. Finally, Meta-SPS grouped the contig containing PRM spectrum 2643 with other overlapping contigs and extracted a meta-contig sequence, which can be seen in Figure S-11A. The resulting meta-contig contained multiple overlapping PRM spectra that can also be seen in Figure S-11A. Figure S-11C shows MS/MS spectrum 2643 annotated with its meta-contig sequence.

References

1. Senko MW, Beu SC, Mclafferty FW (1995) Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry* 6:229-233.

2. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh L-SL (2008) The Universal Protein Resource (UniProt). *Nucleic Acids Research* 35:D190-D195.

3. Horn DM, Zubarev RA, Mclafferty FW (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry* 11:320-332.

4. Zabrouskov V, Senko MW, Du Y, Leduc RD, Kelleher NL (2005) New and automated MSn approaches for top-down identification of modified proteins. *Journal of the American Society for Mass Spectrometry* 16:2027-38.

5. Kullback S, Leibler RA (1951) On Information and Sufficiency. *The Annals of Mathematical Statistics* 22:79-86.

6. Kim S, Mischerikow N, Bandeira N, Navarro JD, Wich L, Mohammed S, Heck AJR, Pevzner PA (2010) The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Molecular & Cellular Proteomics* 9:2840-2852.

7. Bandeira N, Clauser KR, Pevzner PA (2007) Shotgun protein sequencing: assembly of peptide tandem mass spectra from mixtures of modified proteins. *Molecular & Cellular Proteomics* 6:1123-34.

8. Bandeira N, Tsur D, Frank A, Pevzner PA (2007) Protein identification by spectral networks analysis. *Proceedings of the National Academy of Sciences of the United States of America* 104:6140-5.