SUPPL. FIG. 1a

DATA ANALYSIS STRATEGY

Data obtained with the first pass of analysis on tryptic peptides have to be combined with data acquired from 2nd digestion fractions to determine peptide and protein identifications as well as sequence coverages. The data combination can be performed at level of MS/MS spectra prior mascot database search or at the post-interpretation peptide level after database search, with protein or peptide validation software tools like Scaffold.

We compare here the different possible variants.

At MS/MS spectrum level, MS/MS spectra of the large peptide fractions after second digestion were pooled and search against database (**strategy MS1**). The assigned peptides were mapped onto amino acid sequence of corresponding proteins. Comparing the sequence regions mapped by second enzyme data against trypsin data, we could identify the additional sequence coverage.

In another spectrum combination strategy (**strategy MS2**, **strategy 1** of Supplementary figure 1b), the MS/MS spectra of 2nd digestion fractions are pooled with those of all fractions after tryptic digestion and subjected to database search with a mixed enzyme cleavage specificity (e.g. K, R, E for trypsin and Glu-C). In this case, the mascot identified proteins and their corresponding peptides will be listed in one .dat file. It is advantageous in comparison to the strategy MS1 since Scaffold calculates protein probability based on the probabilities of corresponding peptides of the protein and the calculation is solely applied to each (.dat file) MS sample. The more interpreted peptides accumulate in a MS sample the higher the probability of the proteins, resulting in more proteins which could pass the Scaffold filter of 99% protein probability.

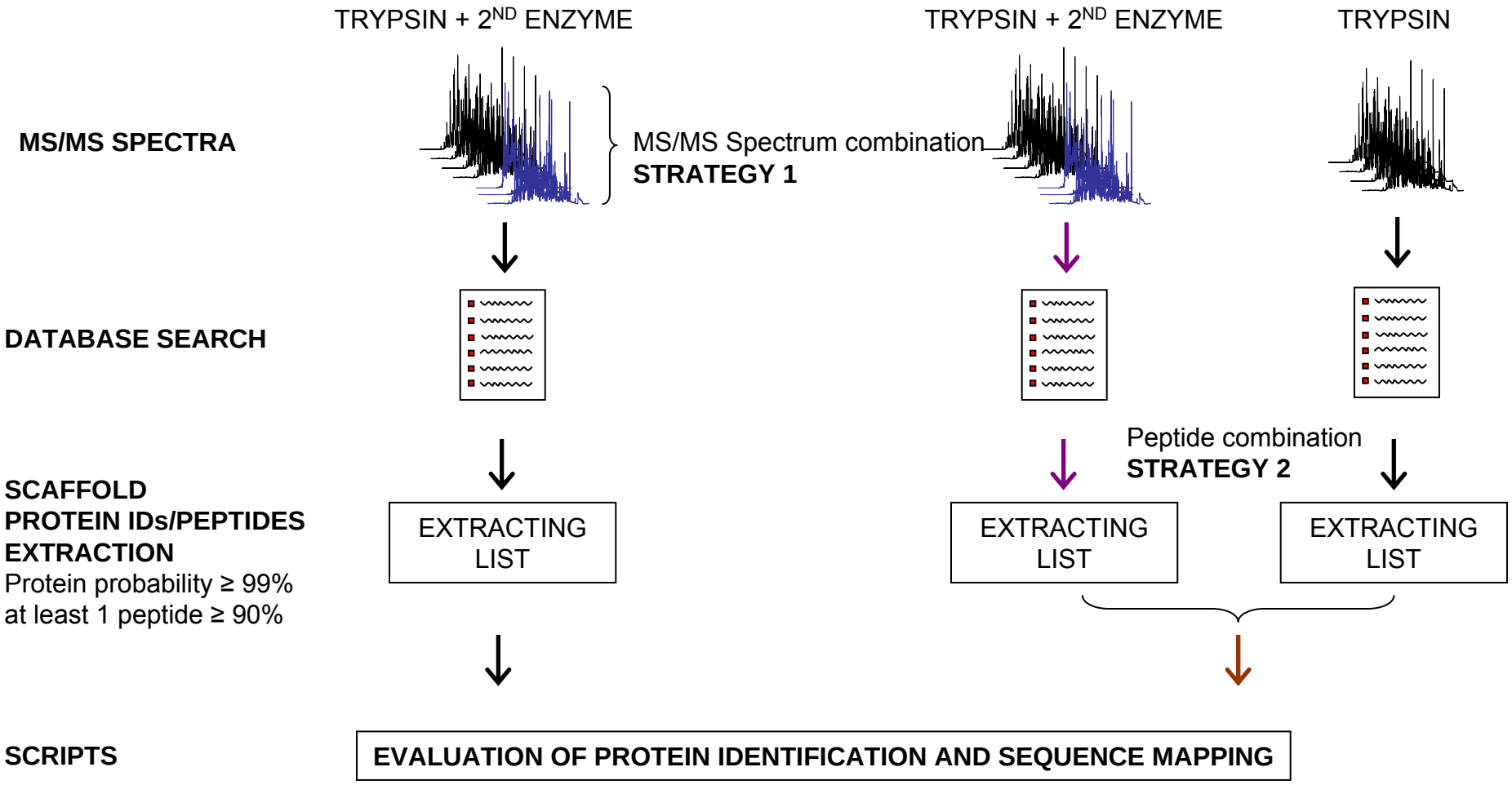
A comparison of the results obtained with the two spectrum combination strategies is displayed in Supplementary Figure 1c. For every cleavage method employed on SolprotSP large peptide SEC fractions, strategy MS2 led to a number of validated proteins 1.6 – 2 times higher, a number of new peps 1.3 – 1.6 times higher and an amino acid sequence coverage 1.3 – 2 times higher than strategy MS1. It means that even though 2nd digestion fractions by themselves provide a number of new-peps, the combination of their MS/MS spectra with tryptic digest spectra before search enhances the number of peptides and proteins passing the Scaffold validation filter because of higher protein probability.

However, alone, the combination of primary and secondary spectrum sets before database search like described for strategy MS2 has a major drawback. In practice, the sequence covered by identified tryptic peptides (peptide groups B and C, see "**Impact of second digestion on the analysis of SEC fractions of a tryptic digest**") dramatically decreased from 20.53% of trypsin alone down to approximately less than 10% in case of strategy MS2 (Supplementary Figure 1c). This is because spectra of pure tryptic peptides containing more than one cleavage site for the second enzyme were not identified because they do not correspond any more to the cleavage criteria. Thus, although some proteins gained an additional 30% sequence coverage, for instance with Glu-C, the average sequence coverage of the identified proteins surprisingly decreases 2 to 6% in comparison with trypsin alone for all investigated samples. Therefore, a merge of all MS/MS spectra of both primary and secondary digests did not provide true sequence coverage values due to the loss of numerous identifications in the trypsin spectra set.

In order to compensate this sequence loss, a post-identification peptide combination procedure (**strategy 2,** Supplementary figure 1b) had to be carried out. The MS/MS spectra of whole tryptic digest were treated separately and with standard parameters from those of second digestion fractions. The set of tryptic peptides obtained was combined with the peptide set of **strategy 1** and then subjected to statistical validation by Scaffold followed by evaluation of protein identification and sequence coverage using custom written scripts. In this case, without any impact of the missed cleavage criteria, all spectra of the tryptic peptides were interpreted adequately.

In brief, two levels of data combination need to be carried out in conjunction to achieve maximal sequence coverage and number of identified proteins. The MS/MS spectra combination of tryptic digest and 2nd digestion fractions is necessary to provide highest number of new-peps and validated proteins while the post-interpretation peptide combination gives correct values for the sequence coverage. An overview of the overall data analysis process is shown in Supplementary figure 1d

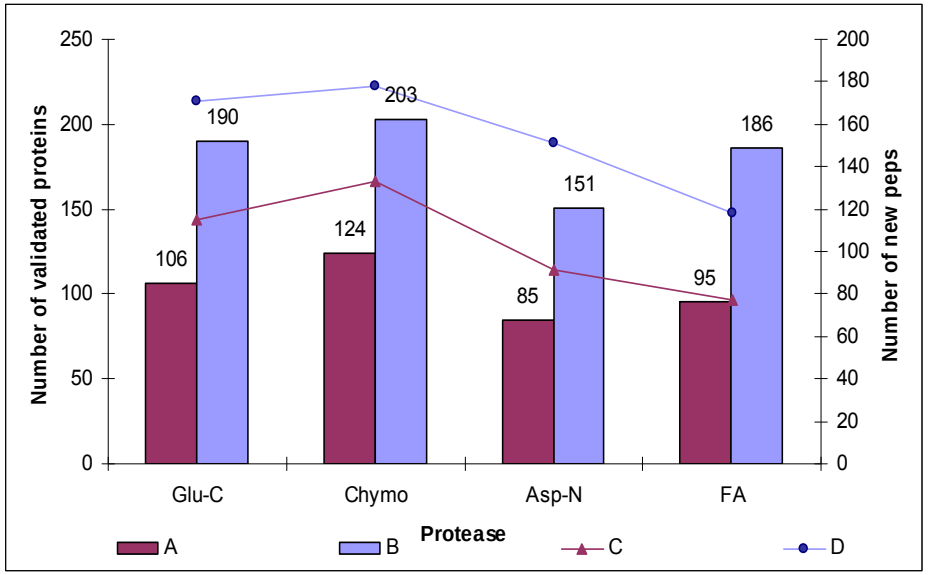
SUPPL. FIG. 1b

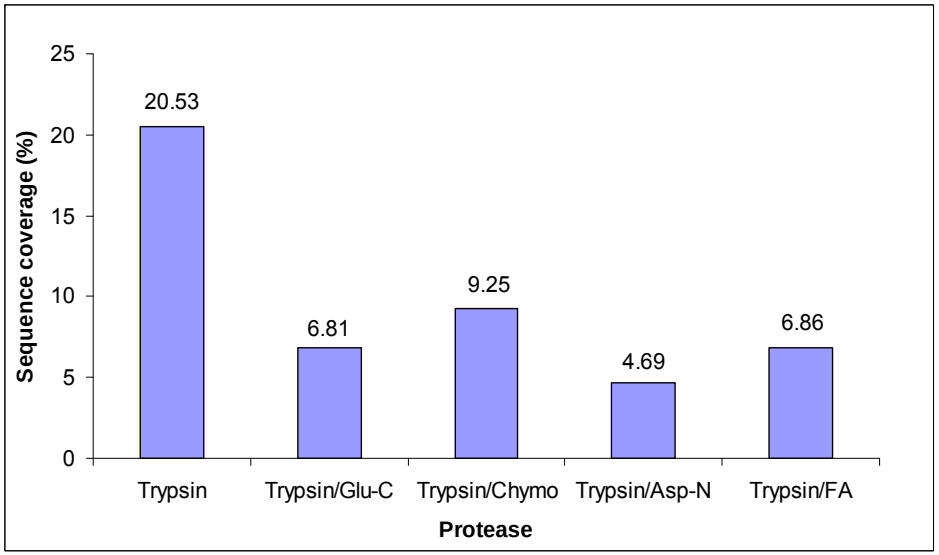


Supplementary Figure 1b:

Strategy 1 (MS2): MS/MS spectrum combination Strategy 2: Post-interpretation peptide combination

SUPPL. FIG. 1c



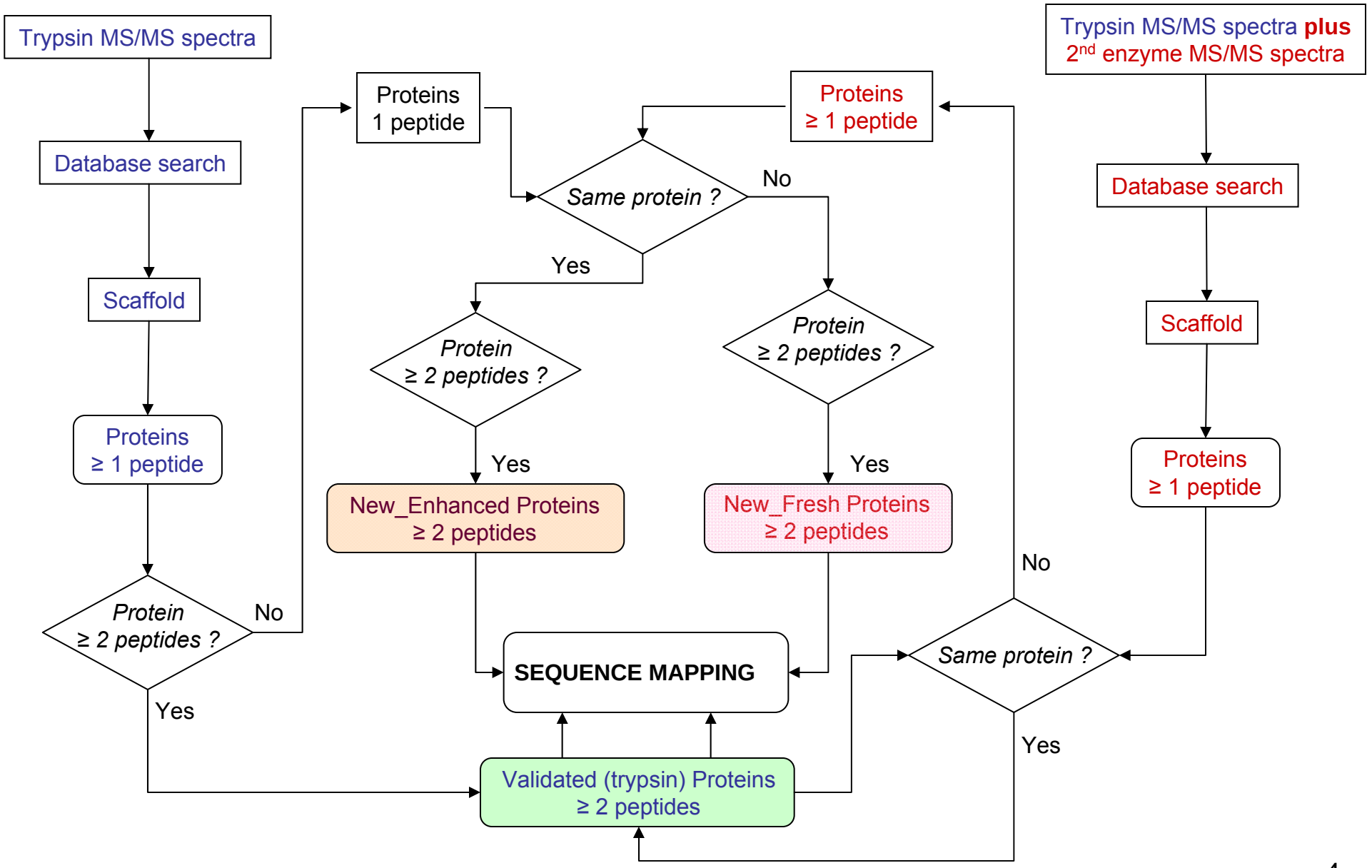


Supplementary Figure 1c-I : Comparison of the numbers of identified proteins (sample solprotSP) obtained from the large peptide SEC fractions after second digestion using the MS/MS spectrum combination strategies MS1 and MS2 (described in Supplementary Figure 1b). Strategy MS2 enhances the number of peptides and proteins passing the Scaffold validation filter, by enhancing protein probability.

A: Number of scaffold validated protein IDs, strategy MS1
B: Number of scaffold validated protein IDs, strategy MS2
C: Number of new-peps, strategy MS1
D: Number of new-peps, strategy MS2

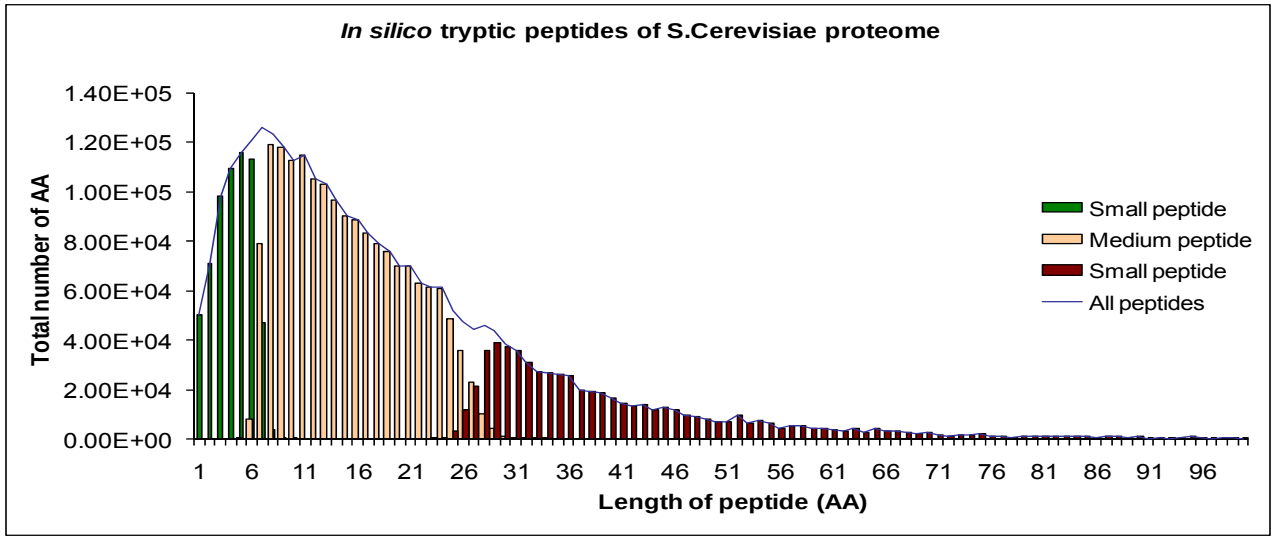
Supplementary Figure 1c-II : Decrease of sequence covered by peptides of tryptic digest when merging MS/MS spectra of tryptic digest and 2nd digestion fractions (strategy 1, strategy MS2), compared to using spectra of tryptic digest alone. The strategy was applied to the data set of 218 identified solprotSP proteins.
The percentage of sequence covered by peptides of tryptic digest decreased from 20.52% to less than 10% when using the strategy MS2. This is due to the inappropriate cleavage specificity used in the MASCOT search which leads to the loss of identifications of peptides containing more than one cleavage site for the secondary cleaving agent.

SUPPL. FIG. 1d

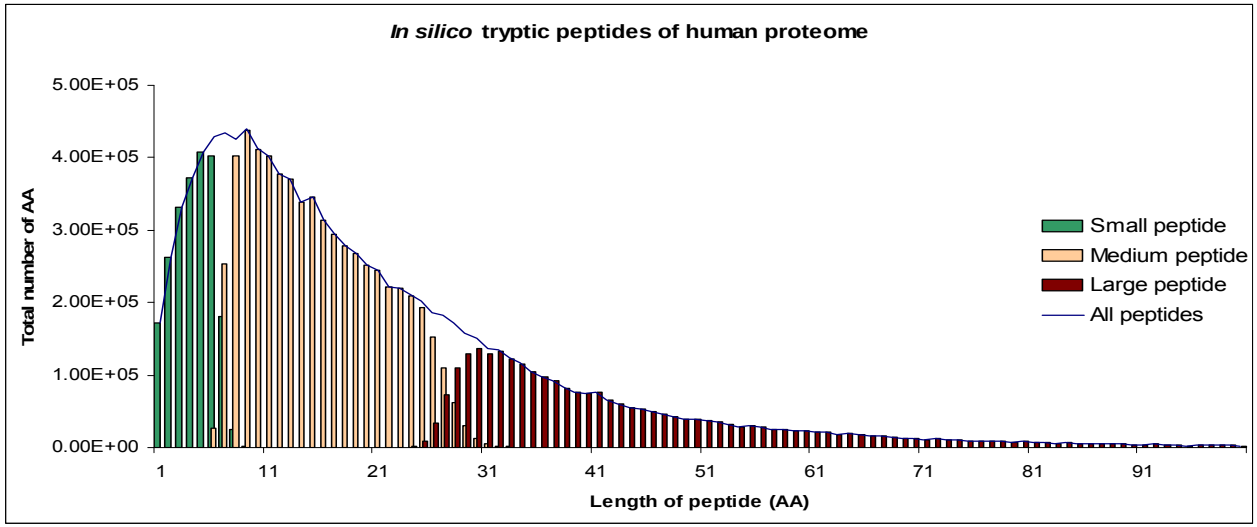


Supplementary Figure 1d: Data processing scheme

SUPPL. FIG. 2

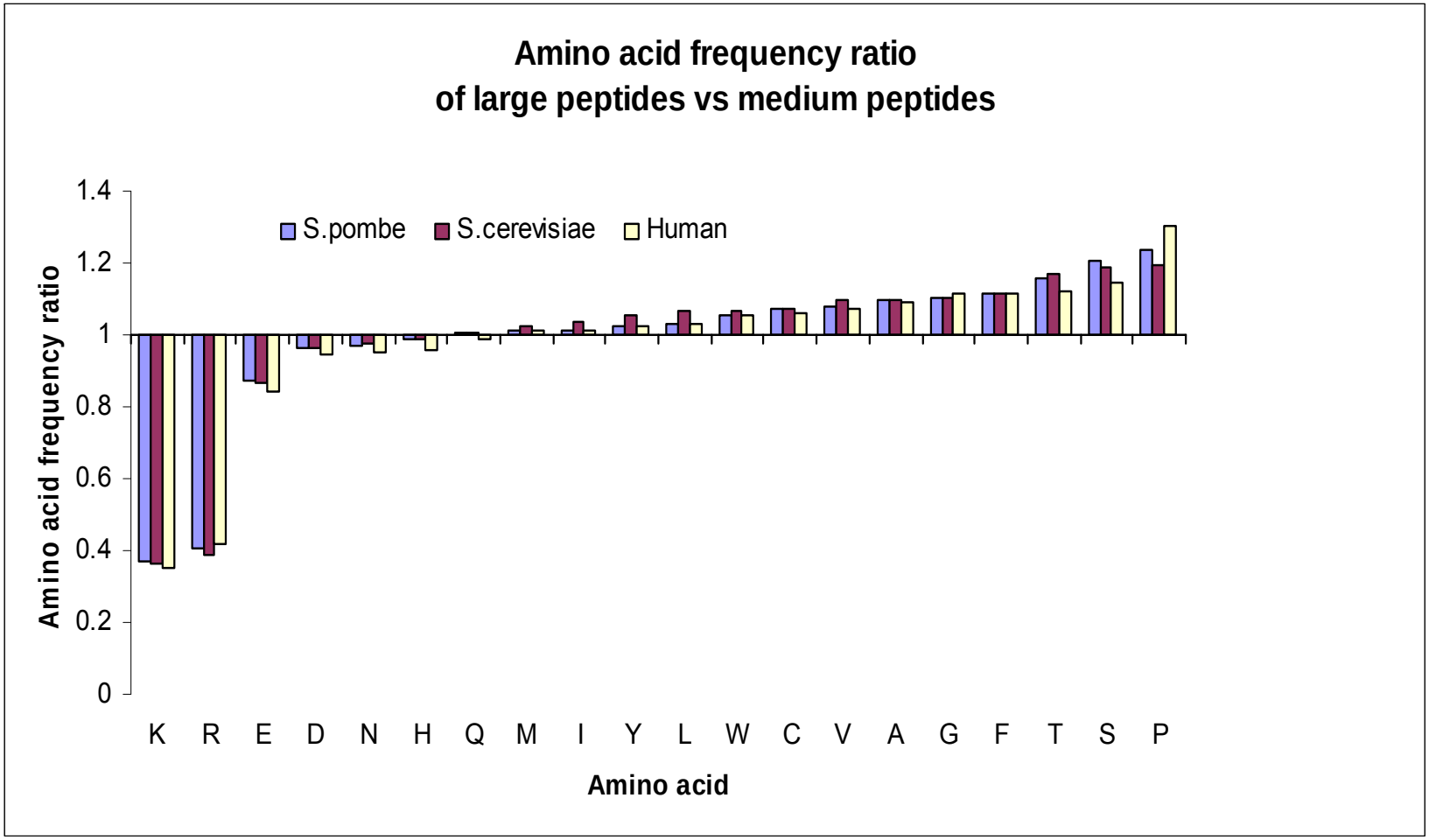


Supplementary Figure 2a: Amino acid content as a function of peptide length (1-100 AA range) for *S. cerevisiae* tryptic peptides, *in silico* generated from 6'552 sequence entries *(*Uniprot/Swissprot database *version 14.8)*. Specificity of trypsin is set to full cleavage.
Large peptide: Mw ≥ 3000 Da, medium peptide: 3000 > Mw > 800 Da and small peptide: M ≥ 800 Da.
The small and medium peptides cover respectively 20.3% and 57.5% of the total proteome. The large peptides covered 22.2%.



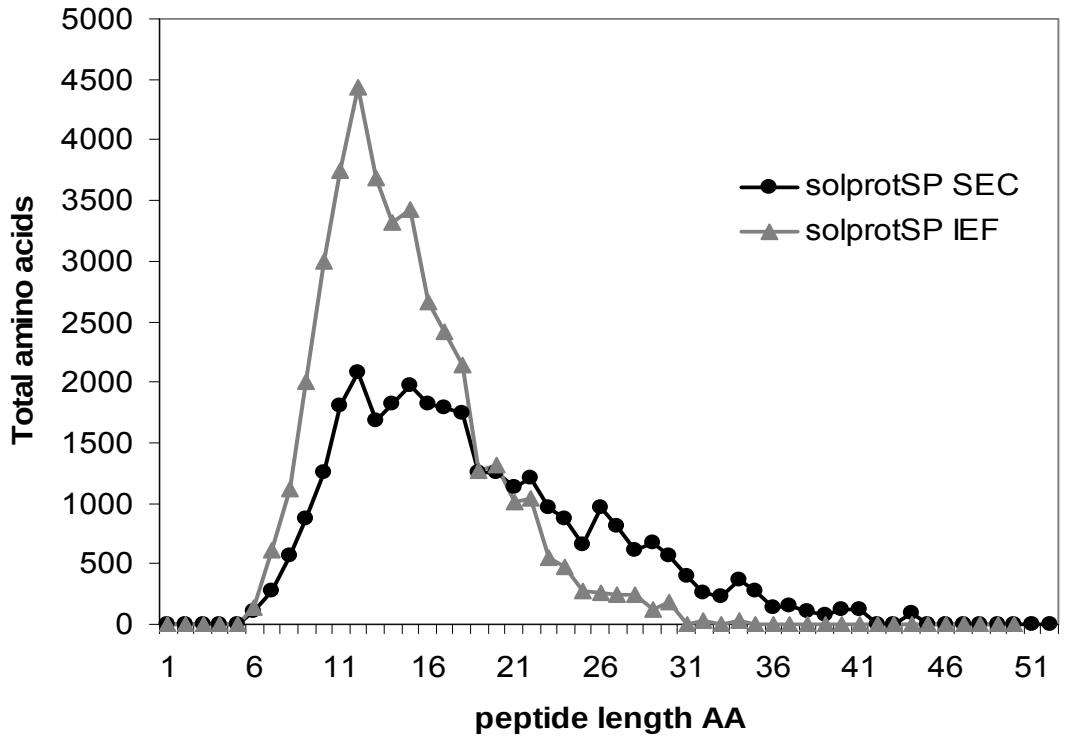
Supplementary Figure 2b: Amino acid content as a function of peptide length for *homo sapiens* tryptic peptides, *in silico* generated from 20266 sequence entries *(*Uniprot/Swissprot database *version 15.15)*. Specificity of trypsin is set to full cleavage.
Large peptide: Mw ≥ 3000 Da, medium peptide: 3000 > Mw > 800 Da and small peptide: M ≥ 800 Da.
The small and medium peptides cover respectively 19.3% and 55.7% of the total proteome. The large peptides covered 25.0%.

SUPPL. FIG. 3



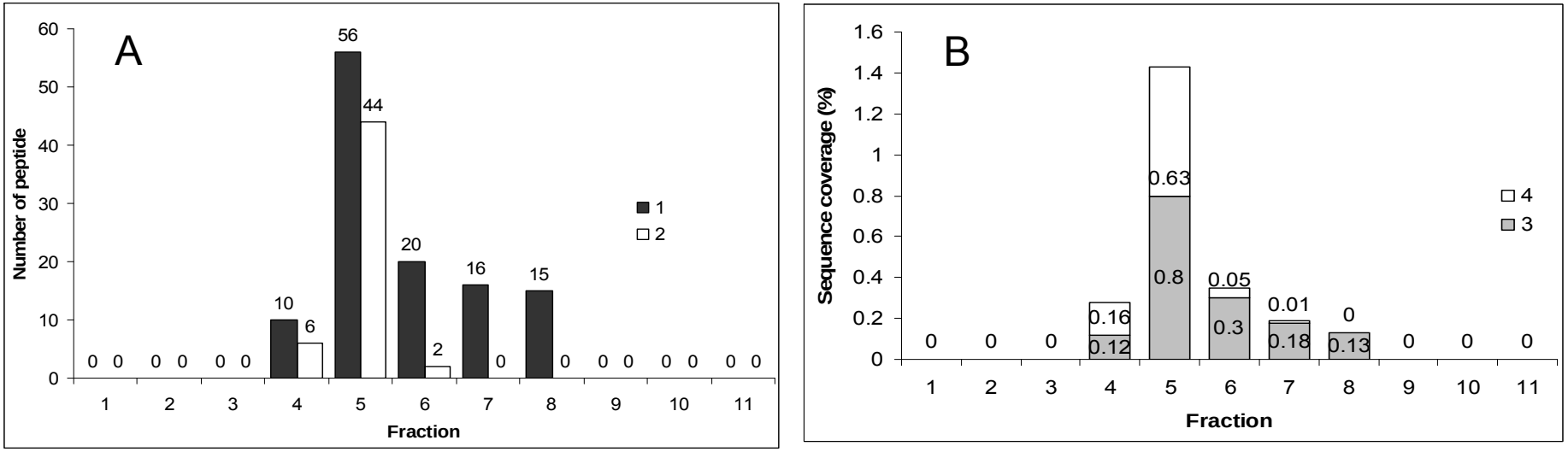
Supplementary Figure 3: Amino acid frequencies as a function of peptide size of large *in silico* tryptic peptides (Mw ≥ 3000Da) to medium *in silico* tryptic peptides (3000 > Mw > 800 Da) for the proteomes of *S.pombe*, *S. cerevisiae* and *H.sapiens*

SUPPL. FIG. 4



Supplementary Figure 4: Amino acid content as a function of length for peptides identified in the tryptic digests of sample solprotSP (S.pombe soluble extract) fractionated either by SEC or IEF.
CID spectra from 11 SEC or 12 IEF fractions were pooled and submitted to database search. Trypsin cleavage was specified with one possible missed cleavage. Data before statistical validation with Scaffold were used for the plot. 9'499 MS/MS spectra identified 1'933 nonredundant peptides which were assigned to 384 proteins for sample solprotSP after SEC (solid black trace), while 15430 spectra identified 683 proteins in the same sample after IEF separation (solid grey trace).

SUPPL. FIG. 5

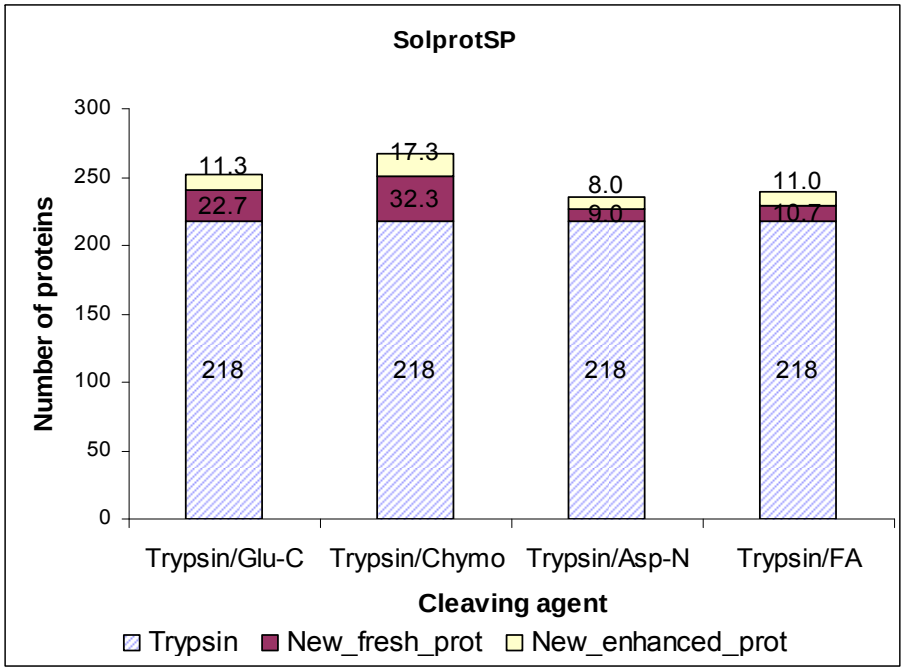


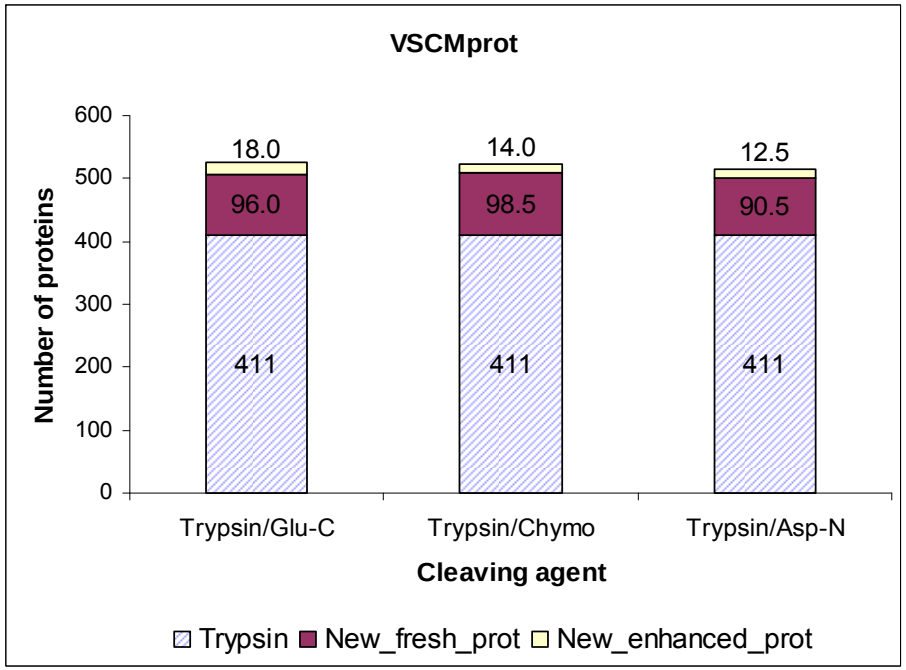
Supplementary Figure 5: Non redundant new peptides (new_peps) identified and increases in sequence coverage after Glu-C secondary digestion as a function of SEC fraction number.

A) Number of new peptides (new-peps) and inferred large parent tryptic peptides (large iPTPs) identified after secondary digestion of the tryptic fractions of the solprotSP tryptic digest with Glu-C. Numbers of non-redundant new-peps are shown, i.e. new-peps found in a fraction were not counted in the following ones. 1: Number of new-peps identified; 2: Number of tryptic peptides with Mr≥ 2'400Da inferred to be precursors of the identified new-peps (i.e. large iPTPs). 117 new-peps were identified in total (sum of all fractions). The total of redundant peptides (group B+C) in the same sample was 504.

B) Percentages of sequence covered by the identified new-peps and their corresponding large iPTPs. 3: Percentage of total sequence covered by new-peps; 3+4: Percentage of total sequence covered by the corresponding iPTPs. The three fractions 4-6 contributed to 86 out of 117 new-peps identified, to 1.22 of 1.53% newly covered sequence and to 52 over 52 extended tryptic peptides with Mr≥ 2.4 kDa.

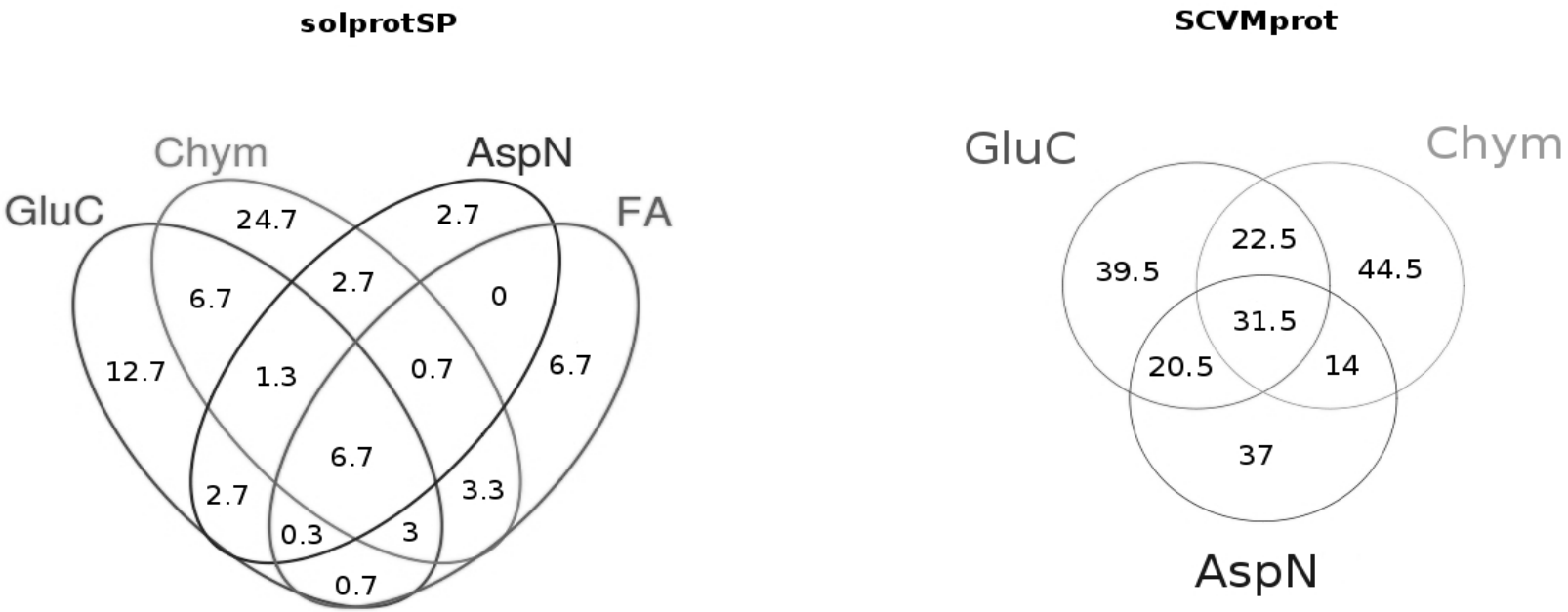
SUPPL. FIG. 6





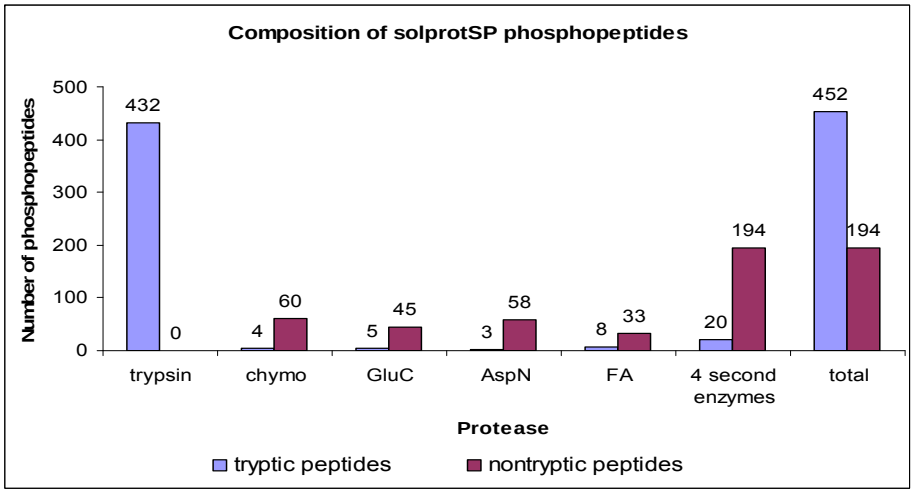
Supplementary figure 6: Number of proteins identified in samples solprotSP and SCVMprot after secondary digestions. Beside the proteins identified in all fractions with trypsin (striped pattern), digesting the large peptide fractions (3 for solprotSP and 4 for SCVMprot) with a 2nd protease or formic acid provided additional new proteins (red and yellow). The values are the average of three and two experimental replicates, respectively. Values are non-redundant relative to trypsin: proteins already found with trypsin alone are not listed. New_fresh_prots are proteins identified by two peptides only found with secondary digestion. New_enhanced_prots are proteins identified by combination of one tryptic peptide with at least one peptide observed after second digestion.
When combining all second digestion results, 75 and 210 new proteins were respectively identified for solprotSP and SCVMprot samples.

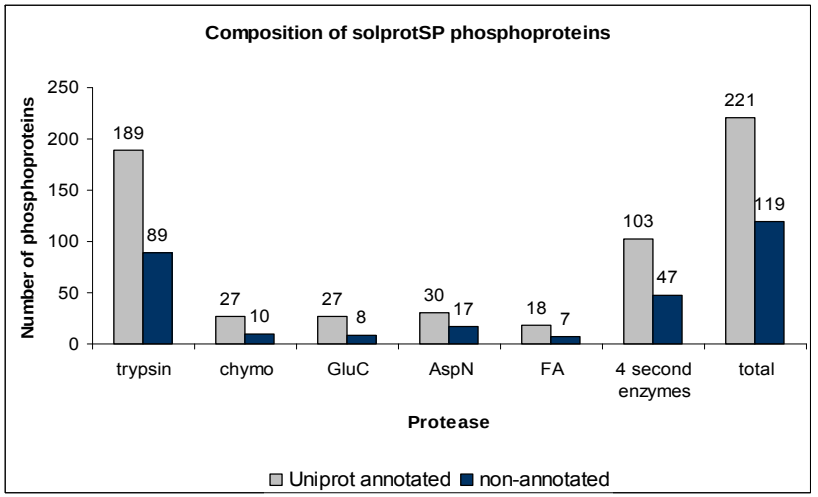
SUPPL. FIG. 7

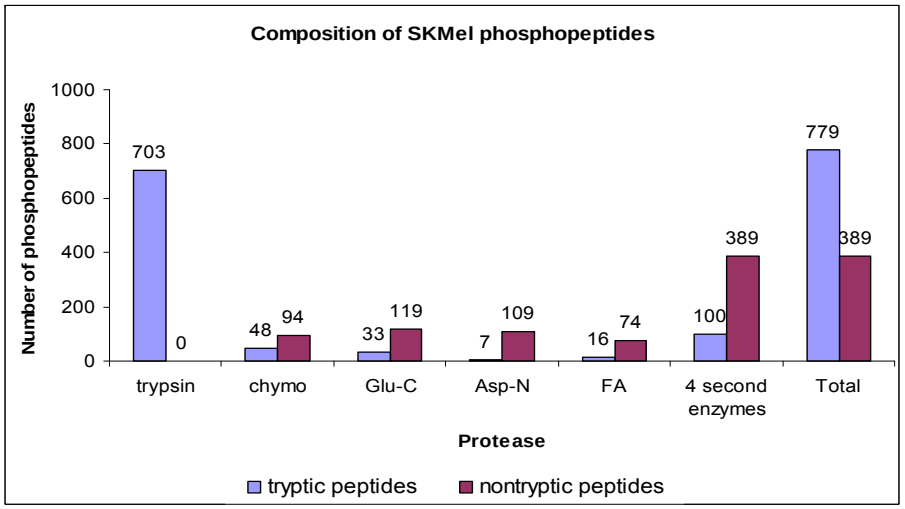


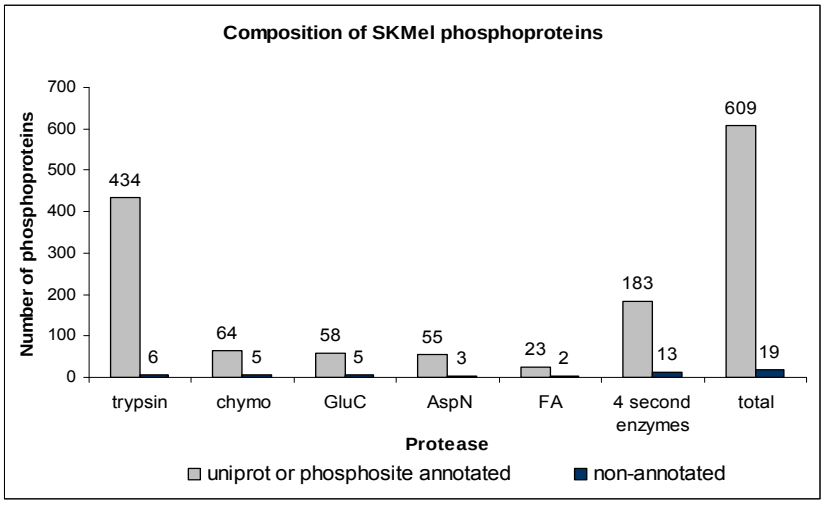
Supplementary Figure 7: New proteins identified in samples solprotSP and SCVMprot after secondary digestion with non-trypsin cleavage agents. Values are average of respectively 3 (solprotSP) and 2 (SCVMprot) experimental replicates.
With the combination of all second proteases, the average numbers of new (not found with trypsin) proteins identified in sample solprotSP (3 technical replicates) and sample SCVMprot (2 technical replicates) were 74.7 and 209.5, respectively. The number of validated trypsin proteins were respectively 218 (solprotSP) and 411 (SCVMprot)

SUPPL. FIG. 8









Supplementary Figure 8 : Numbers of non-redundant phosphopeptides and corresponding phosphoproteins identified with trypsin and other second enzymes or formic acid in samples solprotSP and SKMel.
Phosphopeptides analyses were carried out for all 11 SEC fractions of tryptic digest of solprotSP and SKMel samples. The fractions containing large tryptic peptides (4 for solprotSP and 6 for SKMel) were also analyzed after secondary digestion. An average amount of 100 and 80 µg tryptic peptides respectively of each solprotSP and SKMel SEC fraction was subjected to phosphopeptide enrichment and then a half of it was used for LC-MS/MS analysis.
Numbers given for secondary digestions are non-redundant against trypsin, i.e. if a peptide was already identified in the trypsin dataset it was removed from the list of peptides identified with secondary digestions. Redundancy among secondary digestion datasets was not removed.