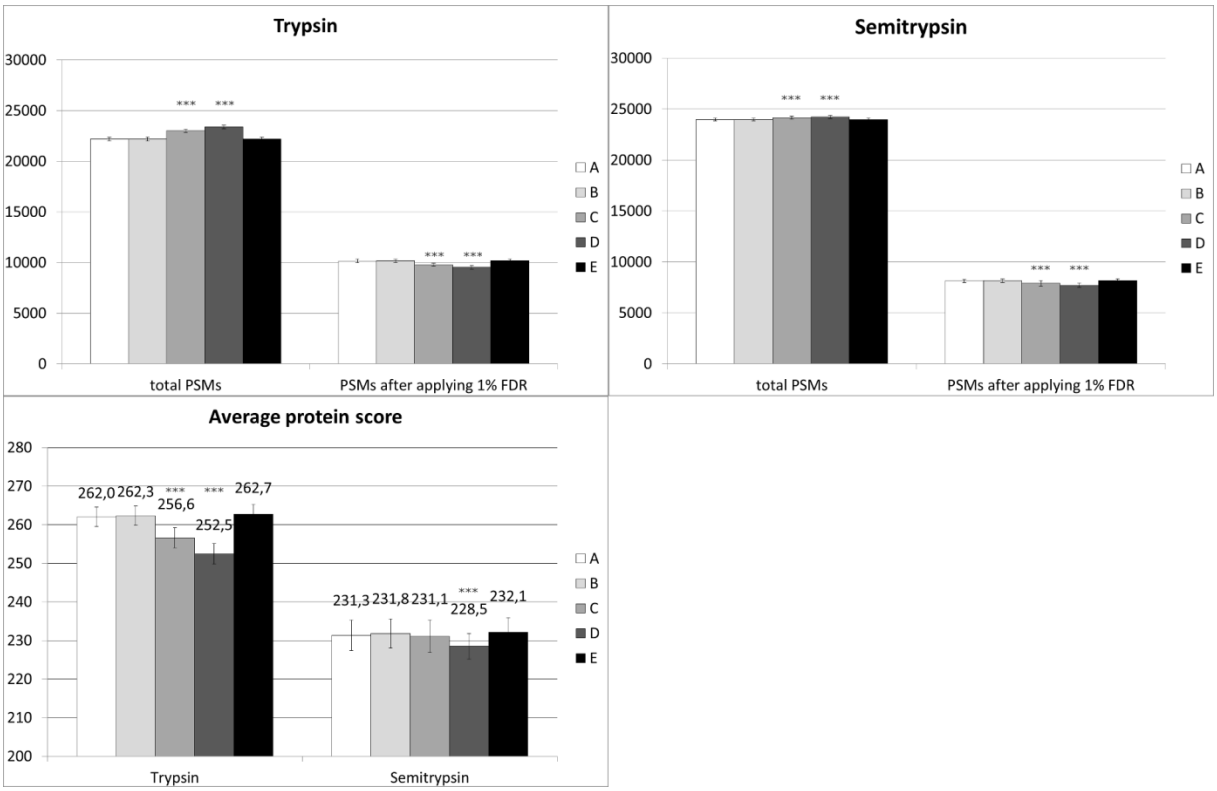
Supporting information

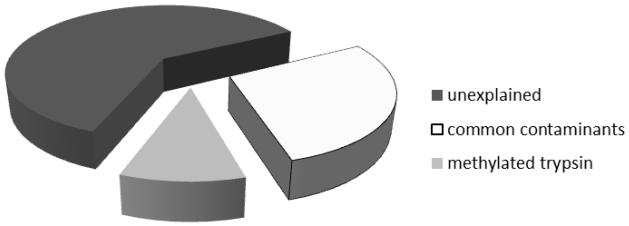
Cleaning out the litterbox of proteomic scientists´ favourite pet: optimised data analysis avoiding trypsin artefacts

Matthias Schittmayer, Katarina Fritz, Laura Liesinger, Johannes Griss, Ruth Birner-Gruenberger



Supporting Figure S1: Impact of different search strategies on total PSM, statistically significant PSM after applying 1% FDR and resulting protein scores. Increased search space results in lower 1% FDR PSMs and lower average protein scores. The effect of semi-trypsin setting is more pronounced than even 2 variable modifications because of the higher number of combinatorial possibilities. Minimizing the search space increase with our artificial amino acid strategy results in more assigned peptides for both, total and 1% FDR PSMs. Databases from Search strategies A-E were reversed using the CompOmics dbToolkit 4.2.3. **A:** Swiss-Prot_Human or Swiss-Prot_Yeast was used as database; carbamidomethylation of cysteines was set as fixed modification. Oxidation of methionine and protein N-terminal acetylation were set as variable modifications. **B:** Swiss-Prot_Human or Swiss-Prot_Yeast with concatenated cRAP was used as database. Carbamidomethylation of cysteines was set as fixed modification. Oxidation of methionine and protein N-terminal acetylation were set as variable modifications. **C**: Swiss-Prot_Human or Swiss-Prot_Yeast with concatenated cRAP was used as database. Carbamidomethylation of cysteines was set as fixed modification.

Oxidation of methionine and protein N-terminal acetylation were set as variable modifications. For dataset 2 acetylation of lysine was allowed as additional variable modification, for all other datasets dimethylation of lysine was allowed as additional variable modification. **D**: Swiss-Prot_Human or Swiss-Prot_Yeast with concatenated cRAP was used as database. Carbamidomethylation of cysteines was set as fixed modification. Oxidation of methionine and protein N-terminal acetylation were set as variable modifications. Methylation and dimethylation of lysine were allowed as additional variable modification. **E:** Swiss-Prot_Human or Swiss-Prot_Yeast with concatenated cRAP was used as database. Furthermore modified trypsin was added to the database (where all lysines were replaced by dimethylated lysines). Carbamidomethylation of cysteines was set as fixed modification. Oxidation of methionine, protein N-terminal acetylation and loss of methylation and loss of dimethylation of dimethylated lysines (J) were set as variable modification. Statistical analysis was done in R using repeated measures ANOVA. Pairwise comparison of different groups (i.e. search strategies A-E (see above) results of 5 biological replicates taken from dataset 3, respectively) was performed using Tukey post-hoc test in R.



Supporting Figure S2: Fractions of false positives explained by different search strategies employing semi-tryptic digest (dataset 3). Search strategy B reveals 28 % of all false positives from search strategy A are caused by common contaminant derived peptides (including unmodified trypsin). Search strategy E identifies additional 10 % which are exclusively caused by methylated trypsin peptides. The fractions of false positives explained by common contaminants and methylated trypsin increases compared to tryptic digest because the total number of hits and therefore false positives is considerably smaller in the semitryptic search (162 vs. 251), the number of spectra matched to common contaminants (46 vs. 58) and methylated trypsin is comparable (16 vs. 17).

Supporting Table S1. Peptides originating from methylated porcine trypsin in a *S. cerevisiae* digest (dataset 1). FDR of 1% was applied. The score and expect value are shown for the best scoring PSM.

Sequence	PSMs	Score	Expect value
R.VATVSLPR.S	97	54	8.0E-06
K.LSSPATLNSR.V	24	63	5.2E-06
K.APVLSDSSCK.S	3	34	1.7E-03
K.APVLSDSSCK.S + Dimethyl (K)	2	23	7.3E-03
K.SSGSSYPSLLQCLK.A	4	52	5.1E-05
K.VCNYVNWIQQTIAAN.-	3	55	7.9E-06
R.SCAAAGTECLISGWGNTK.S	6	107	3.1E-11
R.SCAAAGTECLISGWGNTK.S + Methyl (K)	3	100	2.3E-10
R.SCAAAGTECLISGWGNTK.S + Dimethyl (K)	14	119	3.9E-12
R.LGEHNIDVLEGNEQFINAAK.I	32	126	3.5E-12
R.LGEHNIDVLEGNEQFINAAK.I + Methyl (K)	6	141	1.1E-13
R.LGEHNIDVLEGNEQFINAAK.I + Dimethyl (K)	10	92	4.9E-09
K.IITHPNFNGNTLDNDIMLIK.L	4	92	1.0E-08
K.IITHPNFNGNTLDNDIMLIK.L + Methyl (K)	1	48	2.2E-04
K.IITHPNFNGNTLDNDIMLIK.L + Oxidation (M)	3	79	1.3E-07
K.IITHPNFNGNTLDNDIMLIK.L + Methyl (K); Oxidation (M)	1	38	2.0E-03
K.SSGSSYPSLLQCLKAPVLSDSSCK.S + Dimethyl (K)	1	43	4.8E-04
R.IQVRLGEHNIDVLEGNEQFINAAK.I	1	85	4.7E-08
K.IITHPNFNGNTLDNDIMLIKLSSPATLNSR.V	1	69	1.4E-06
K.IITHPNFNGNTLDNDIMLIKLSSPATLNSR.V + Methyl (K)	1	51	7.7E-05
K.IITHPNFNGNTLDNDIMLIKLSSPATLNSR.V + Dimethyl (K)	1	60	5.7E-06

Supporting Table S2: Peptides originating from a Promega modified trypsin self-digest. FDR of 1% was applied. The score and expect value are shown for the best scoring PSM.

Sequence	PSMs	Score	Expect value
K.LSSPATLNSR.V	1	16	2.6E-02
R.LGEHNIDVLEGNEQFINAAK.I	13	42	6.6E-05
R.LGEHNIDVLEGNEQFINAAK.I + Methyl (K)	5	54	4.3E-06
R.LGEHNIDVLEGNEQFINAAK.I + Dimethyl (K)	17	33	5.4E-04
K.IITHPNFNGNTLDNDIMLIK.L	3	34	4.1E-04
K.IITHPNFNGNTLDNDIMLIK.L + Oxidation (M)	7	23	5.2E-03
K.IITHPNFNGNTLDNDIMLIK.L + Dimethyl (K)	1	19	1.1E-02
K.IITHPNFNGNTLDNDIMLIK.L + Methyl (K); Oxidation (M)	2	14	3.9E-02
K.IITHPNFNGNTLDNDIMLIK.L + Dimethyl (K); Oxidation (M)	1	21	8.9E-03
R.IQVRLGEHNIDVLEGNEQFINAAK.I	1	48	1.4E-05
K.IITHPNFNGNTLDNDIMLIKLSSPATLNSR.V	2	27	1.9E-03
K.IITHPNFNGNTLDNDIMLIKLSSPATLNSR.V + Dimethyl (K)	1	16	2.8E-02

Supporting Table S3: Peptides originating from acetylated trypsin in dataset 2. Unmodified and acetylated peptides of trypsin could be identified when allowing acetylation of lysine as variable modification.

Peptide number	Sequence	PSMs	Score	Expect value
1	K.SAASLNSR.V	2	49	1.4E-05
2	K.LKSAASLNSR.V	1	15	3.2E-02
3	K.LKSAASLNSR.V + Acetyl (K)	3	29	1.3E-03
4	K.SSGTSYPDVLK.C	1	70	9.6E-08
5	K.VCNYVSWIK.Q	2	42	6.8E-05
6	K.LQGIVSWGSGCAQK.N	2	128	1.7E-13
7	K.VCNYVSWIKQTIASN.- + Acetyl (K)	1	33	5.4E-04
8	R.LGEDNINVVEGNEQFISASK.S	19	153	4.7E-16
9	K.SIVHPSYNSNTLNNDIMLIK.L + Oxidation (M)	1	77	1.8E-08

Supporting Table S4: Summary of PRIDE cluster search results. Result of the PRIDE cluster analysis of all current human datasets (209) for wrongly assigned spectra originating from modified trypsin. (See separate file SupportingTableS4.xls)