Supplement to: "Triqler for Data Independent Aquisition Data"

Patrick Truong

Matthew The

Lukas Käll

October 19, 2021

Note S1: xxx

Comparison of ability to differentiate differentially abundant proteins

Constant variance

	Unfiltered	no_shared	no_shared_IL_equivalence
All	31 055	30 456	30 452
E. Coli	4 391	4 306	4 306
Human	20 614	20 302	20 299
Yeast	6 050	5 848	5 847

Table S1: **Protein count in the Uniprot FASTA protein database.** The database is a FASTA file with one protein sequence per gene for each species (UP000005640, UP000000625 and UP000002311. Acquired on 2021-06-16). The "no_shared" filter is applied by splitting the protein sequences at amino acids "K" and "R" and keeping all the sequences with length > 7 for each protein. We then mapped each sequence to all possible protein matchings and counted the how many proteins each split sequence was mapped to, and filtered so that each sequence only kept one protein match. Therefore, creating a database library with only one peptide sequence per protein. For the "no_shared_IL_equivalence" all I are replaced by L before performing the filtering.

Comparison of statistical calibration

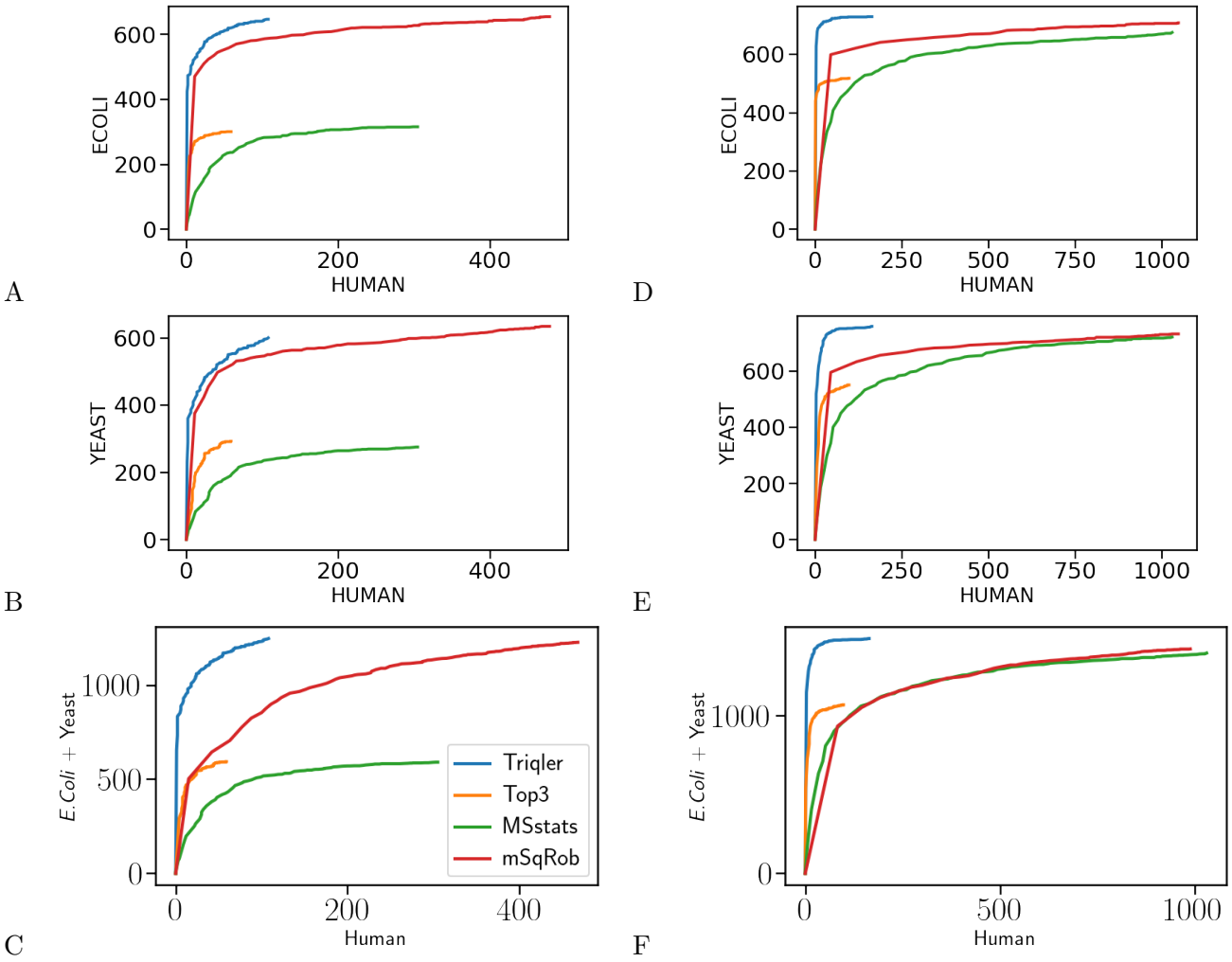

Figure S1: **Comparison of ability to differentiate differentially abundant proteins** We plotted the number of reported differentially abundant *E. Coli* and Yeast proteins as a function of number of proteins from the HeLa background when sorting according to significance for (A) DDA generated spectral libraries and (B) DIA-Umpire geneated Pseudo spectra. For the test we selected a fold-change treshold of 0.4 for Triqler.

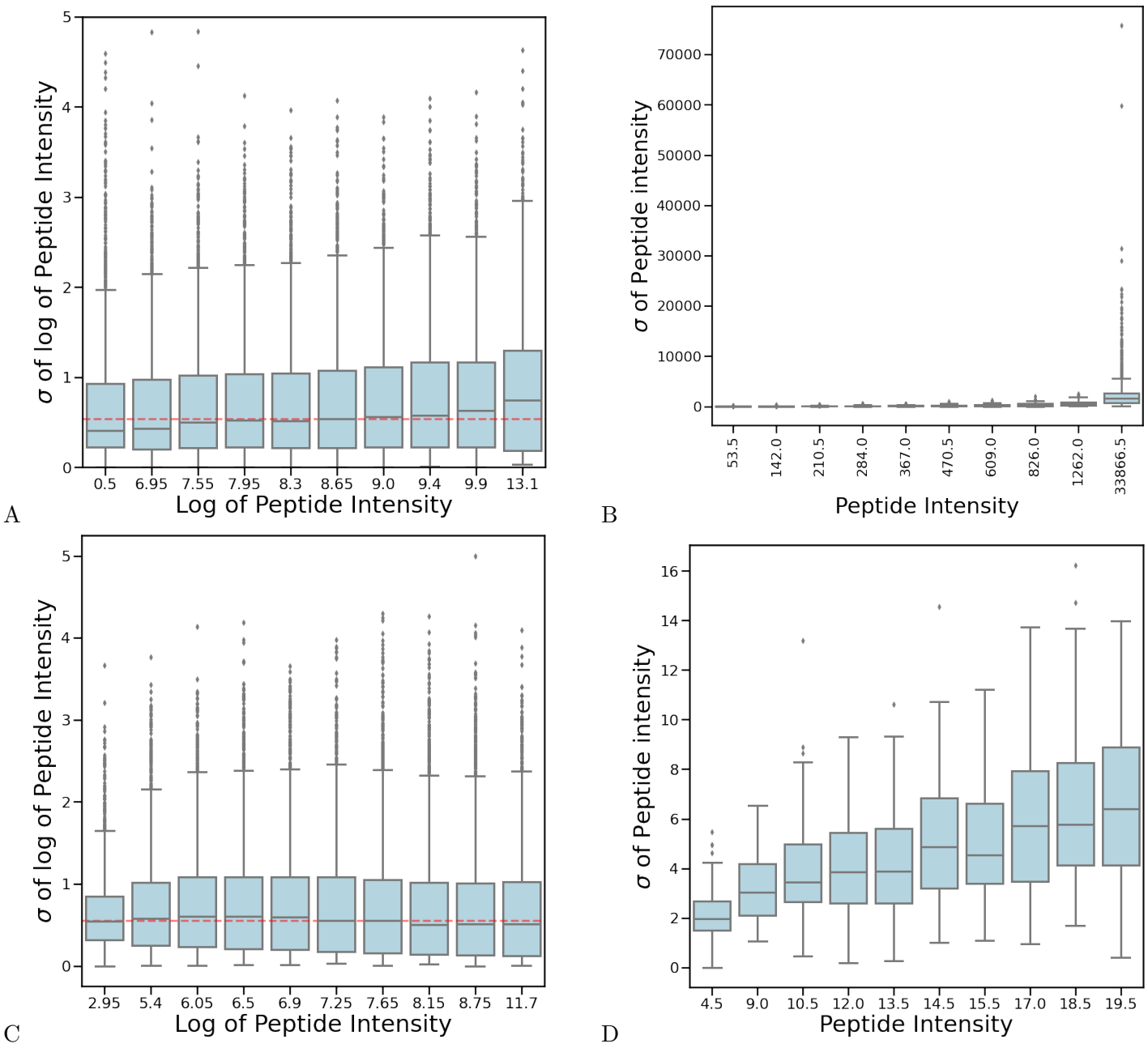

Figure S2: **Uniform offset in standard deviation.** We plotted the standard deviation as a function of the mean of every peptide intensity in the TripleTOF6600 section of the LFQ Bench set in a linear and log-log scale for (A-B) spectral library matching and (C-D) pseudo-spectra workflows. We observe a nearly uniform offset in standard deviation across the intensity scale, demonstrating that $\log(\sigma) \approx \log(\mu) + \log(k)$ and hence $\sigma \approx \mu k$. Linear scale plots are reported as reference.

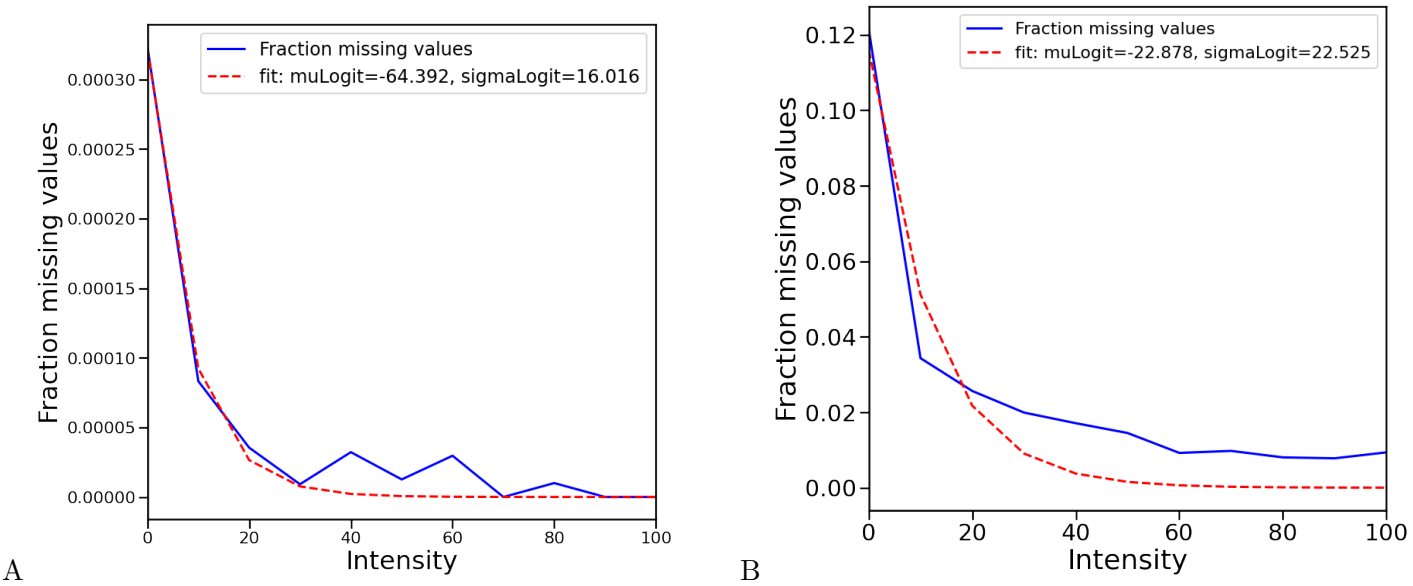

Figure S3: **Comparison of actual missing values against fit to the censored normal distribution used in triqler [?].** We imputed the missing values as the mean of sample peptide intensities and used these imputed values to approximate the missingness for a given intensity. We binned the intensities to an arbitrary small range and plotted the fraction os missing values for each intensity range. A cubic spline fit was used to fit the values against the mentioned censored normal distribution for (A) Spectral library matching and (B) Pseudo spectra workflow.