Supplemental Data

Supplemental Data to "Improved Normalization of Systematic Biases Affecting Ion Current Measurements in Label-free Proteomics Data"

Paul A. Rudnick[1*], Xia Wang[2*], Xinjian Yan[1], Nell Sedransk[3] and Stephen E. Stein[1]

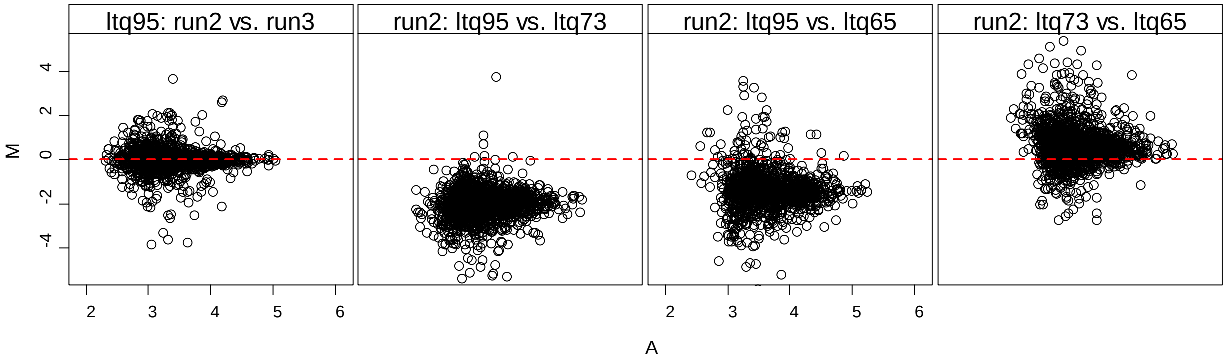
[1]Biomolecular Measurement Division, National Institute of Standards and Technology, 100 Bureau Dr., MS 8362, Gaithersburg, MD; [2]Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH; [3]National Institute of Statistical Sciences, Research Triangle Park, NC

* These authors contributed equally to this work.

Contents

Figure S1 Systematic bias in ion current measurements measured by the relative intensities (M) and its relationship with selected variables for the experimental runs on the LTQ instruments in Study 8. The relative intensity M is defined as the log2 ratio, $M = \log2 (I_{R1}/I_{R2})$, where I_{R1} and I_{R2} are the intensities for the experimental run R1 and R2, respectively. The selected variables included the absolute abundance $A = 0.5 [\log10 (I_{R1}) + \log10 (I_{R2})]$, precursor m/z, $z/\sqrt{peptide\ length}$, and retention time (RT).

S1A: The relative intensity (M) vs. abundance (A) within and across instruments. All experimental runs are 300 ng/µl yeast samples ('high'). *Panel 1*: ltq95 (2nd run) vs. ltq95 (3rd run); *panel 2:* ltq95 (2nd run) vs. ltq73 (2nd run); *panel 3*: ltq95 (2nd run) vs. ltq65 (2nd run); *panel 4:* ltq73 (2nd run) vs. ltq65 (2nd run).



S1B: Boxplots of the relative intensities (M) under the 4 observed charge states (+1, +2, +3, +4) on the ltq73 instrument in Study 8. The 1st and 2nd runs of 300 ng/µl yeast sample ('high') were used in the pair for technical replicates. The 1st run of 60 ng/µl yeast sample ('low') and the 2nd run of 300 ng/µl yeast sample('high') were used in the pair of 5-fold difference. M values were grouped by charge states. Significant difference in M values between charge states was common when samples are differently loaded (5-fold difference). The distribution similarity across charge states was tested by a two-sample Wilcoxon rank test. As expected, the distributions between the charge states across samples were statistically different (p <0.05) with the exception of +1 compared to +2 (p=0.81) and +1 compared to +3 (p=0.23) for the pair with 5-fold difference. Surprisingly, for the technical replicates in this example, the distribution of the doubly charged were significantly different than that of the triply charged (p <0.001). All other comparisons were not. This may be due to the fact that these M calculations were based on raw, un-normalized, abundance values. Together, these results indicated that precursor charge state is an important variable to be considered during data normalization.

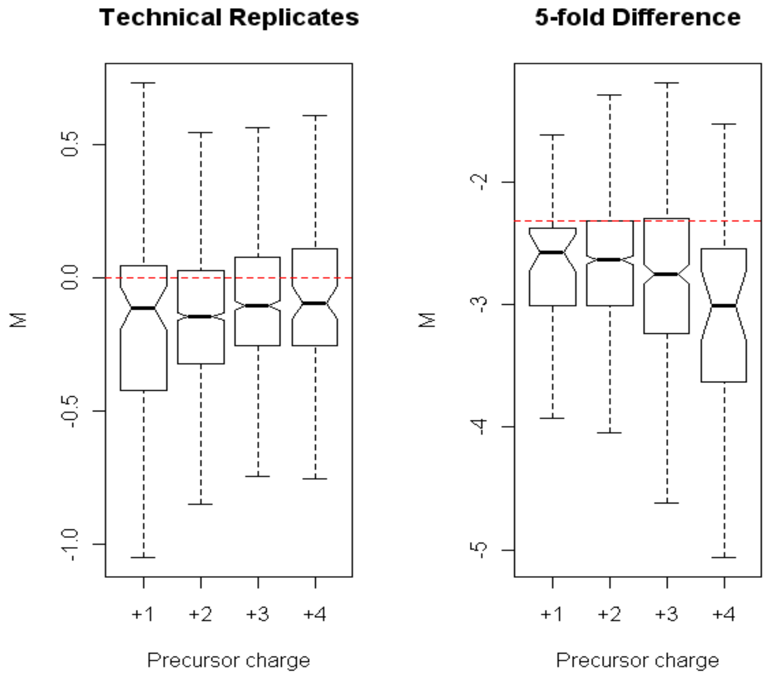


Figure S2 Median relative abundance deviations in the relative intensities (M) versus retention time (RT) quartiles. All experimental runs are from Orbitrap 65 in Study 8. The technical replicates pair includes the 2nd and the 3rd runs in the 300 ng/µl yeast samples ('high'). The 5-fold difference pair includes the 2nd run in the 300 ng/µl yeast sample ('high') and the 2nd run in the 60 ng/µl yeast sample ('low'). The solid line shows the technical replicates pair and the dashed line shows the 5-fold difference pair. This plot revealed large RT biases in the fourth quartile (Q4) when intensities between different samples were compared.

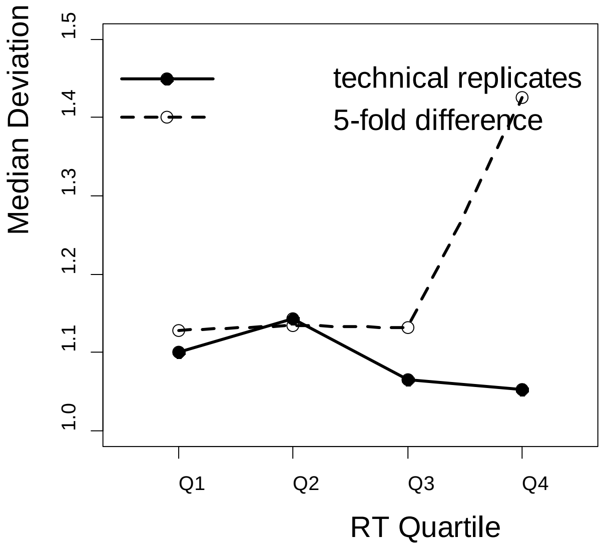


Figure S3 The flow chart of the normalization and variable ranking algorithm.

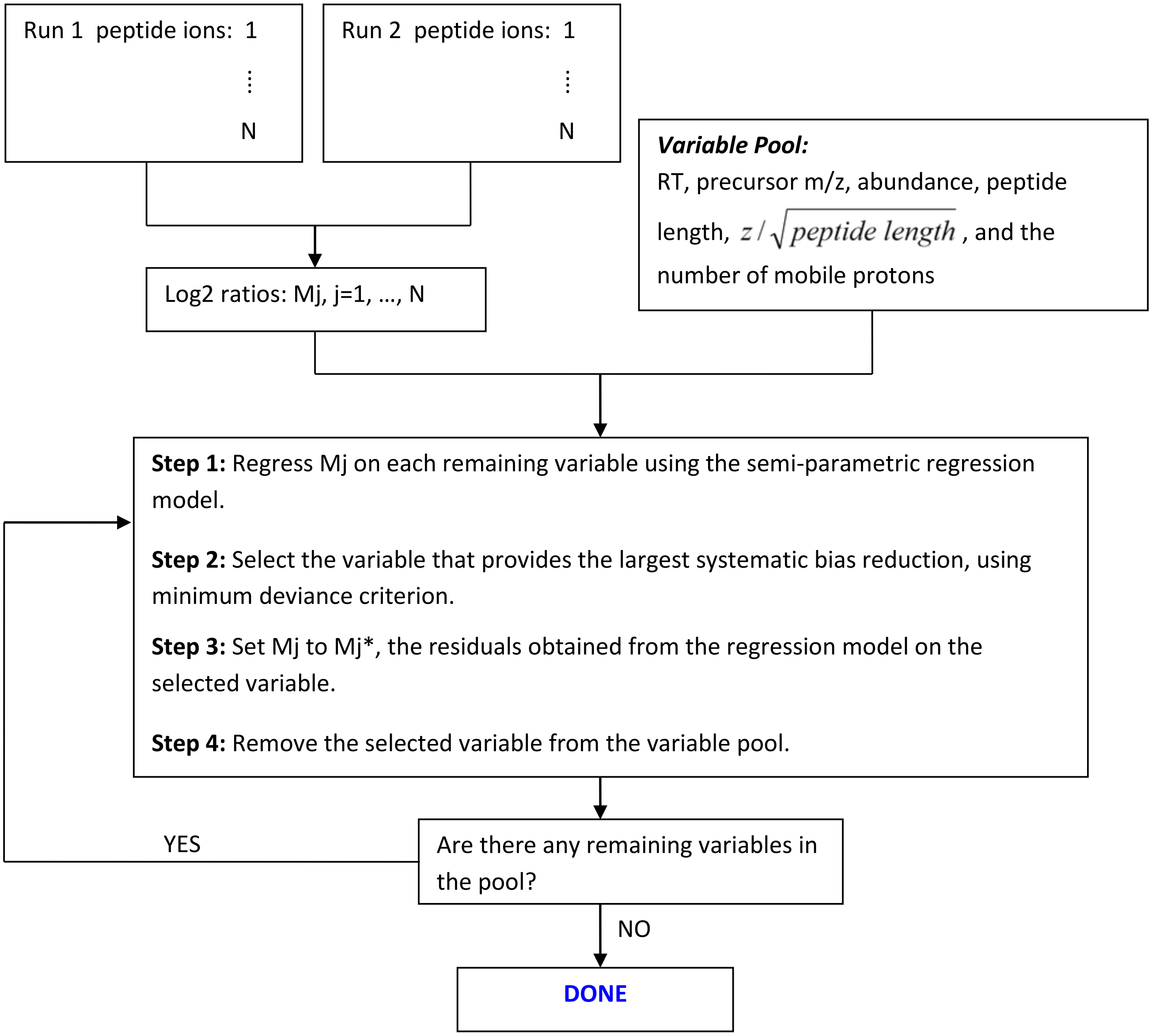

Run 1 peptide ions: 1 ⋮ N
Run 2 peptide ions: 1 ⋮ N
Log2 ratios: Mj, j=1, …, N
Variable Pool: RT, precursor m/z, abundance, peptide length, $z/\sqrt{peptide\ length}$, and the number of mobile protons
Step 1: Regress Mj on each remaining variable using the semi-parametric regression model.
Step 2: Select the variable that provides the largest systematic bias reduction, using minimum deviance criterion.
Step 3: Set Mj to Mj*, the residuals obtained from the regression model on the selected variable.
Step 4: Remove the selected variable from the variable pool.
Are there any remaining variables in the pool?
YES
NO
DONE

Figure S4 Algorithm demonstration. The data used were the 1st run of 300 ng/µl yeast sample ('high') on the ltq95 instrument and the 2nd run of 60 ng/µl yeast sample ('low') on the ltq73 instrument in Study 8. A total of 6 variables were regressed on M in order of decreasing importance (i.e., rank). The six variables included were retention time (RT), precursor m/z, abundance, peptide length, $z/\sqrt{peptide\ length}$, and the number of mobile protons. The relative intensity is defined as M = log2 (I_{R1}/I_{R2}), where I_{R1} and I_{R2} are the intensities for the experimental run R1 and R2, respectively. The solid lines are the fitted regression curves. The adjusted M's are the residuals from the regression of M on the given variable and D is the deviance. The variable giving the smallest deviance is ranked highest. In this example, RT was the rank 1 variable (circled in Step 1). The adjusted M values from step 1 were the residuals from the regression of M on RT. Proceeding to step 2, RT was removed from the variable pool, and the adjusted M from step 1 was regressed on the remaining variables to select the next variable that gave the smallest deviance. The adjusted M values from step 2 were obtained as the residuals from the regression of M on the selected variable (A in this example). The iterative process continues until all variables have been ranked. The M values from the last step become the final normalized M values.

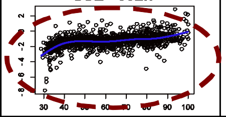
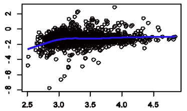
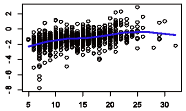
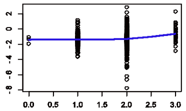
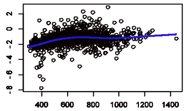
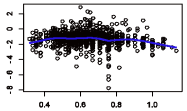
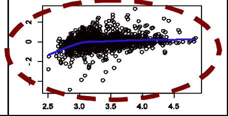
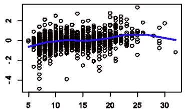
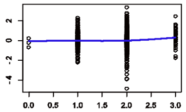
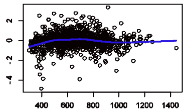
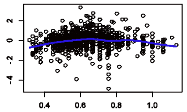
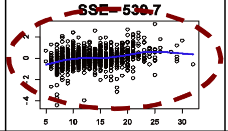
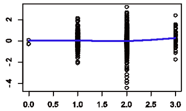
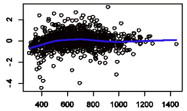
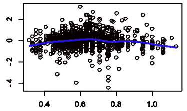
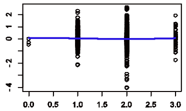
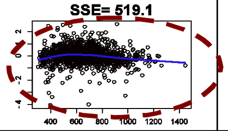
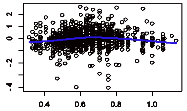
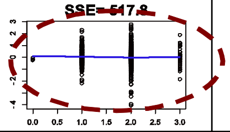
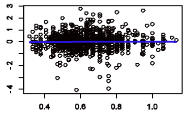
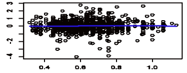
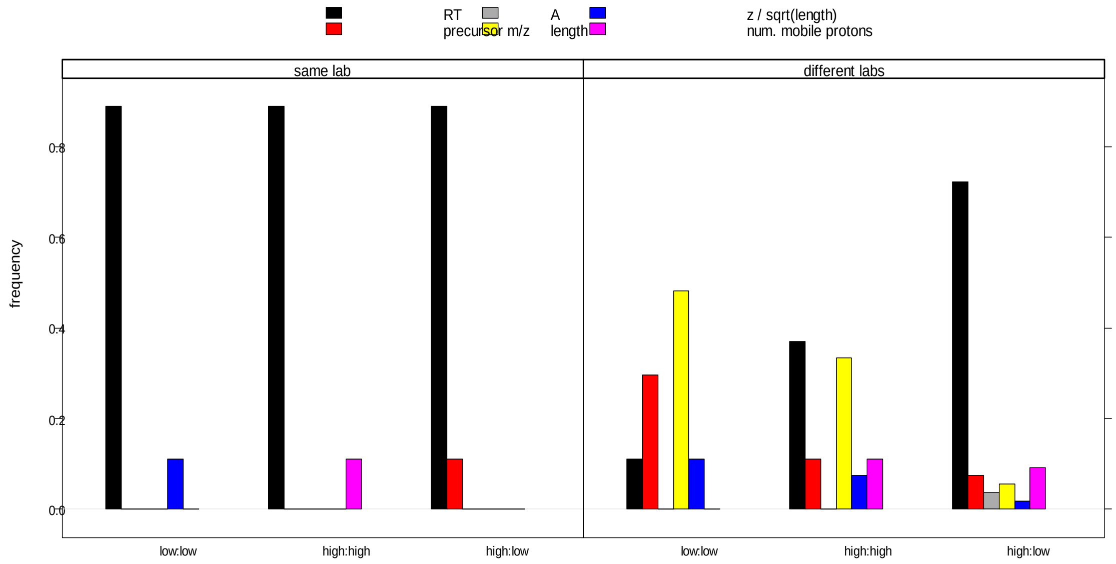
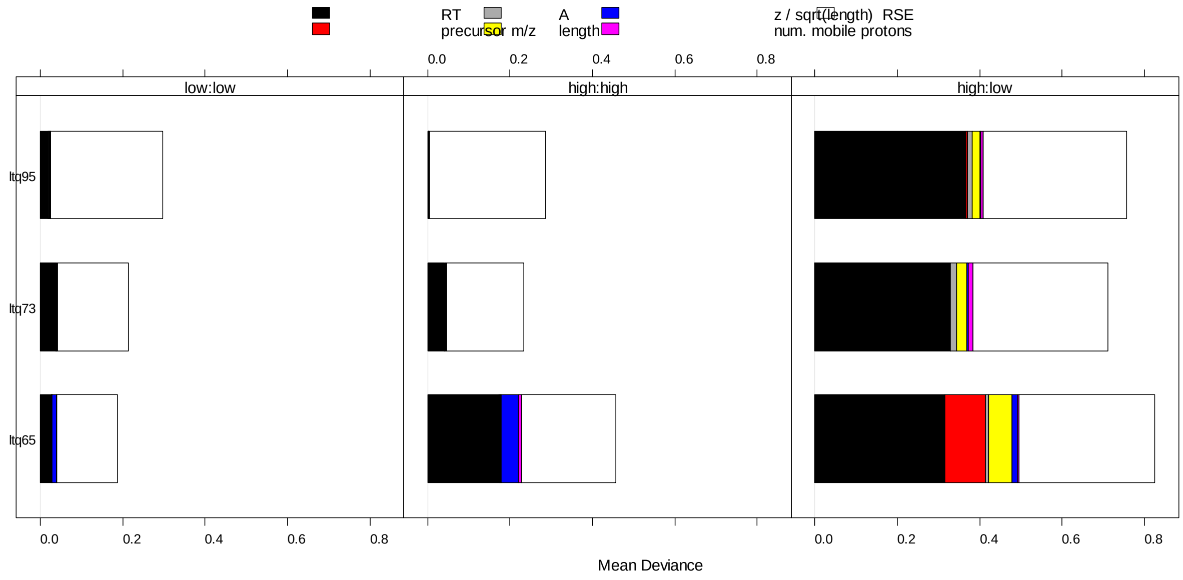
	RT	A	Peptide length	Mobile protons	Precursor m/z	$z / \sqrt{\text{peptide length}}$,
1	SSE= 652.7 	SSE= 820.5 	SSE= 717.9 	SSE= 855.4 	SSE= 767.8 	SSE= 829.8 
2		SSE= 592 	SSE= 601.7 	SSE= 647.9 	SSE= 618.1 	SSE= 618.5 
3			SSE= 530.7 	SSE= 587.4 	SSE= 553.8 	SSE= 567.7 
4				SSE= 538.1 	SSE= 519.1 	SSE= 524.4 
5				SSE= 517.8 		SSE= 518.6 
6						SSE= 517 

Figure S5 The ranking and mean deviances of the variables in normalization. The data used were 18 experimental runs on the 3 LTQ instruments from Study 8, including 60 ng/µl (‘low’) and 300 ng/µl (‘high’) yeast samples.

S5A: The frequency of the variables as Rank 1 for runs within the same lab or across different labs.



S5B: The magnitude of the mean deviance adjusted by each variables as well as the remaining mean deviance (represented by RSE) when experimental runs were from the same labs.



S5C: The magnitude of the mean deviance adjusted by each variables as well as the remaining mean deviance (represented by RSE) when experimental runs were from different labs.

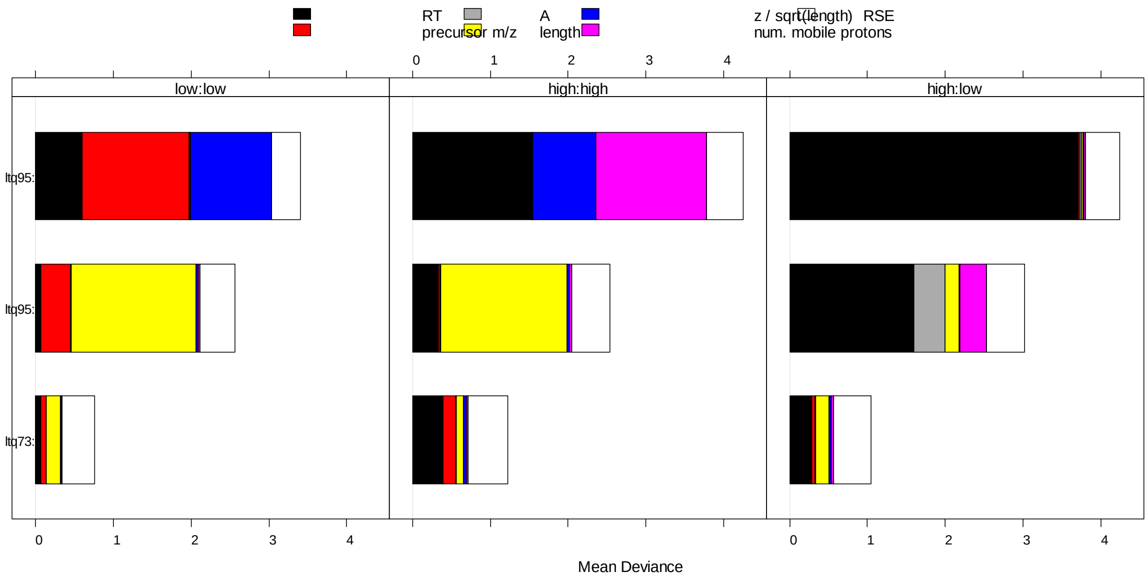
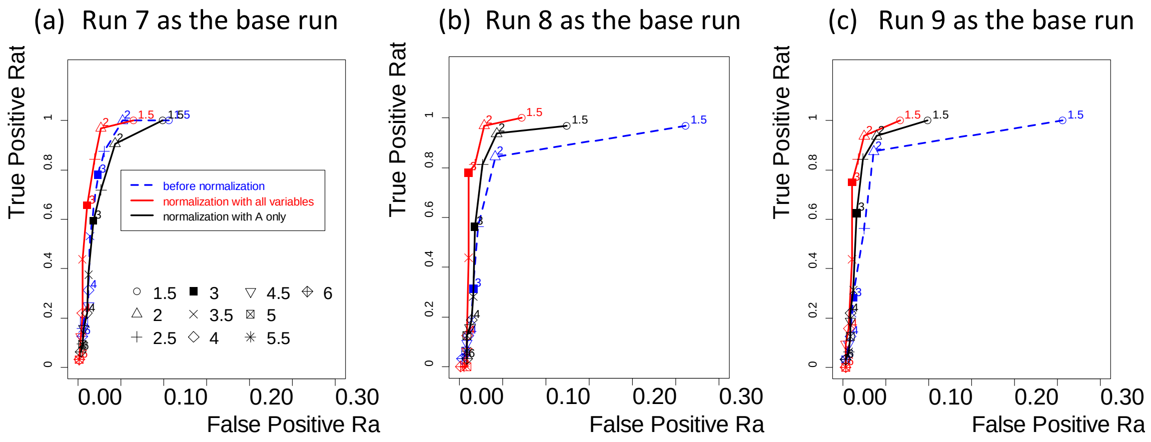


Figure S6 The ROC curves before normalization, normalized by A only and normalized by all variables for Sample 6C (yeast + UPS1 at 2.2 fmol/µl) against Sample 6D (yeast + UPS1 at 6.7 fmol/µl) from the first lab in Study 6. Each of the runs from the low concentration (Sample 6C, run #7, 8, 9) is used in the numerator (base run) to calculate the relative intensities (M) in a pair with the runs from the high concentration (Sample 6D, run #10, 11,12).

S6A: Normalization used the global rank-invariant set (18) as the set of common peptide ions.


(a) Run 7 as the base run (b) Run 8 as the base run (c) Run 9 as the base run

S6B: Normalization used yeast peptide ions as the set of common peptide ions, which was known in Study 6.

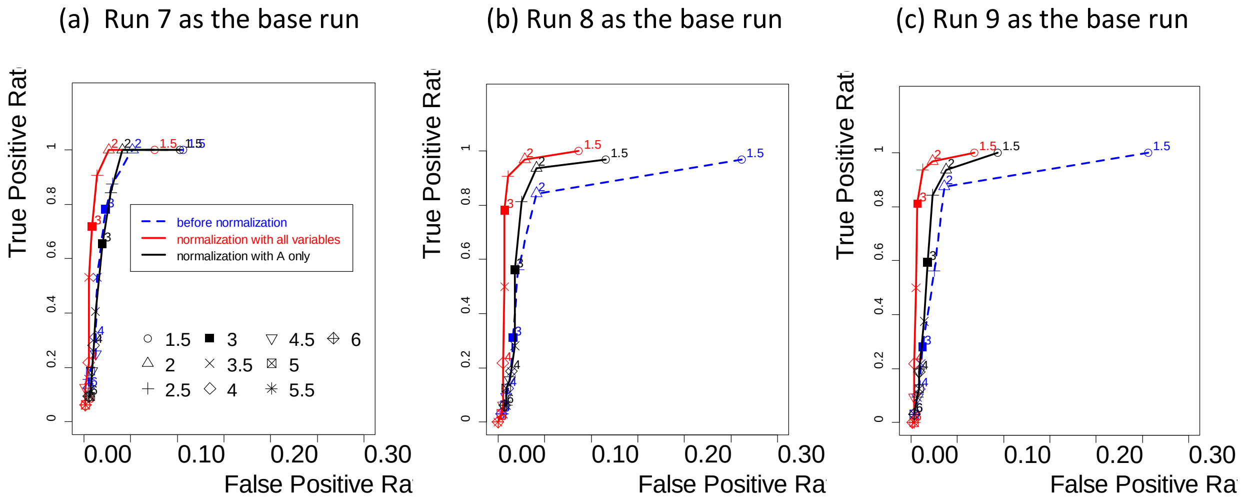

(a) Run 7 as the base run (b) Run 8 as the base run (c) Run 9 as the base run

Table S1: The sensitivity and false positive rate (FPR = 1-specificity) with the 3-fold decision criterion for Sample 6C (yeast + UPS1 at 2.2 fmol/µl) against Sample 6D (yeast + UPS1 at 6.7 fmol/µl) from the first lab in Study 6. Each of the runs from Sample 6C (run #7, 8, 9) was used in the numerator (base run) to calculate the relative intensities (M) in a pair with the runs from Sample 6D (run #10, 11, 12). The results in Column 'before' used data before normalization. The results in Column 'With A only' used data normalized by the abundance (A) only. The results in Column 'With all variables' used data normalized by all variables. The normalization used the yeast peptide ions as the set of common peptide ions, which was known in Study 6.

	Sensitivity			FPR		
	Before	With A only	With all variables	Before	With A only	With all variables
base run=7	0.78	0.66	0.72	0.023	0.020	0.009
base run=8	0.31	0.56	0.78	0.016	0.018	0.007
base run=9	0.28	0.59	0.81	0.013	0.018	0.007