SUPPLEMENTARY MATERIAL TO PEPA test: fast and powerful differential analysis from relative quantitative proteomics data using shared peptides

Laurent Jacob[1] Florence Combes[2] Thomas Burger[2]

[1]Université de Lyon; Université Lyon 1; CNRS; UMR 5558,
Laboratoire de Biométrie et Biologie Évolutive, Lyon, France

[2]BIG-BGE (Université Grenoble Alpes, CNRS, CEA, INSERM),
38000 Grenoble, France

laurent.jacob@univ-lyon1.fr

June 30, 2017

Abstract

Section A and B provide preliminary comparisons between the various peptide to protein aggregation methods: Section A focuses on classical aggregation methods while Section B is dedicated to SCAMPI, that has been proposed to account for shared peptides in isotope labelling proteomic experiments. Section C provides preliminary comparisons between the classical tests used to detect differentially abundant proteins. Finally, we propose additional plots in Section D. Supplementary material to "More powerful differential analysis of relative quantitative proteomics", by Jacob *and others*.

A Preliminary comparisons: aggregation

Aggregation methods can be differenciated according to two criteria: first, the involved operator (sum or mean of peptide intensities, possibly followed by normalization according to each protein properties, see for instance Silva *and others* (2006)) and second, the set of retained peptides. The first one is an important question in absolute quantitative proteomics but has little influence in relative quantification: any difference between these operators equally applies to both compared conditions, so that at protein level, the significance of the differential abundance is not affected.

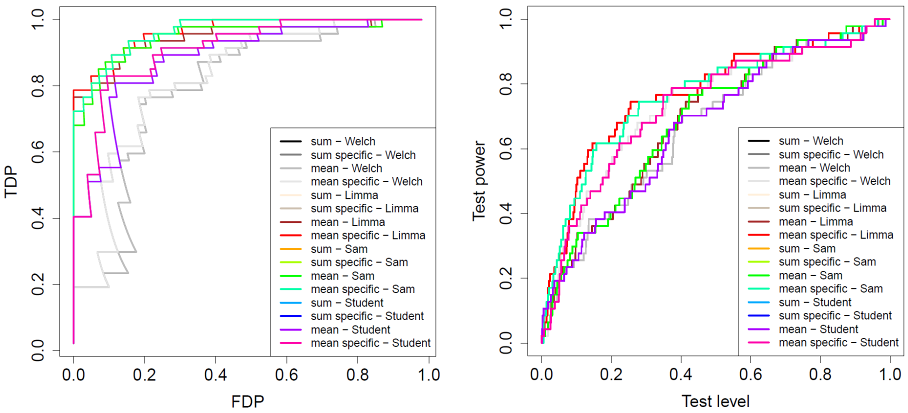
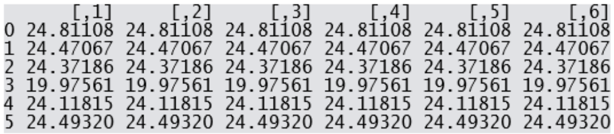


Figure 1: Performance plots of the various aggregation methods on a real (left) and simulated (right) dataset. Curves related to sum-based aggregation are systematically hidden by curves related to mean-based aggregation, for they lead to exactly equal results. Whatever the statistical test, using all peptides rather than only the protein-specific ones leads to a decrement of the performances.

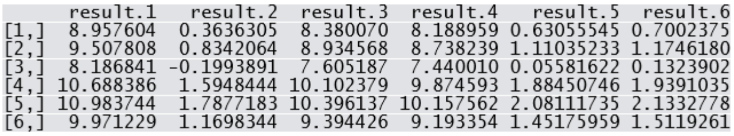
With regard to the second point, it appears (see Figure 1) that using all peptides indistinctly (*i.e.* both specific and shared peptides are considered as if they were protein-specific) leads to less accurate differential analysis than only relying on specific ones, as shared peptides generally leads to protein abundance overestimation. Similarly, using only the most abundant peptides is less efficient as it leads to a loss of information. This last point is more thoroughly discussed in the experimental section of the article.

B Preliminary comparisons: SCAMPI

SCAMPI (Gerster *and others*, 2014) was initially proposed for absolute quantification experiments with isotope labelling, to estimate the ratio of the labelled protein over the original one for each protein within a given sample. The latter can then be inferred on the basis of the known concentration of the former, that has been artificially introduced in the sample. It is thus tempting to apply SCAMPI statistical framework to relative quantification. To do so, SCAMPI authors suggest "*running SCAMPI on each replicate/condition separately*". However, our experiments showed this procedure is not accurate. To explain this, we have considered a dummy dataset where the same sample was replicated 6 times, so as to mimic a perfect experiment with no instrumental variability. We have then applied SCAMPI to each replicate separately as suggested. We expected equal protein abundance estimates, but obtained rather different results as illustrated on Figure 2: although within-sample differences of abundance seem to be respected (as for isotope/original protein abundance ratios), between-sample protein abundances are not.



```
         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
0 24.81108 24.81108 24.81108 24.81108 24.81108 24.81108
1 24.47067 24.47067 24.47067 24.47067 24.47067 24.47067
2 24.37186 24.37186 24.37186 24.37186 24.37186 24.37186
3 19.97561 19.97561 19.97561 19.97561 19.97561 19.97561
4 24.11815 24.11815 24.11815 24.11815 24.11815 24.11815
5 24.49320 24.49320 24.49320 24.49320 24.49320 24.49320
```



```
      result.1   result.2  result.3  result.4   result.5  result.6
[1,]  8.957604  0.3636305  8.380070  8.188959 0.63055545 0.7002375
[2,]  9.507808  0.8342064  8.934568  8.738239 1.11035233 1.1746180
[3,]  8.186841 -0.1993891  7.605187  7.440010 0.05581622 0.1323902
[4,] 10.688386  1.5948444 10.102379  9.874593 1.88450746 1.9391035
[5,] 10.983744  1.7877183 10.396137 10.157562 2.08111735 2.1332778
[6,]  9.971229  1.1698344  9.394426  9.193354 1.45175959 1.5119261
```

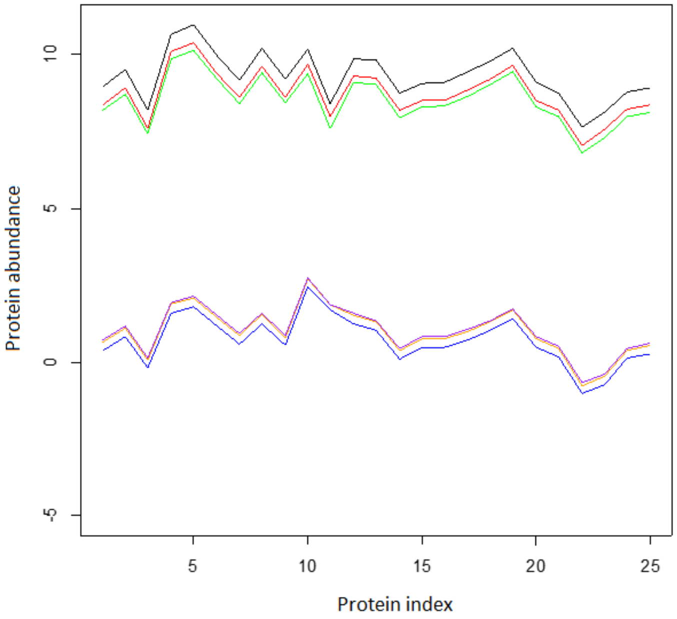


Figure 2: Illustration of SCAMPI results on relative quantification experiments: Top, the six first lines of a fictive dataset with exactly equal replicates; Middle, the corresponding results as provided by the iterative application of SCAMPI on the six replicates; Bottom, the abundance display of the first 25 proteins of each sample (each colored line correspond to a specific sample). Contrarily to what is expected, there are significant differences between the supposedly equal samples.

C Preliminary comparisons: testing procedures

As with aggregation, several methods are reported in the proteomics literature to select putative differentially abundant proteins. The oldest one is based on selecting the proteins with the greatest fold-change (Ting *and others*, 2009), that is the absolute value of the difference (between the conditions) of mean log-transformed abundances. Although practically efficient, this method is nowadays hardly used, as it does not allow to control the false discovery rate associated to the set of selected proteins. Statistical tests are now classically considered, and followed by multiple test corrections. While specific tests are required for spectral count data (such as for instance the Beta-Binomial test Pham *and others* (2010)), extracted ion chromatogram data fit well the assumptions underlying the t-test. Several variations are classically considered:

- The original Student t-test and its Welch generalization to conditions with different numbers of replicates;
- SAM Tusher *and others* (2001), where the variance estimate is regularized by a fudge factor;
- Limma Smyth (2005), where the variance estimates are shrinked across proteins.

We report the results of an experiment comparing these procedures in Figure 1.

Student outperforms Welch The Welch t-test is theoretically of interest to process datasets with missing values. In label-free proteomics experiments however, there are often too few replicates per condition to deal with missing values, and imputation must be conducted first (Lazar *and others*, 2016), so that the interest of Welch t-test is disputable. In our experiments with equal number of replicates within each condition and after imputation we observe that Student's t-test systematically outperforms Welch's. This is probably caused by the small number of samples which makes separate variance estimation per group more difficult.

Regularization helps As expected from the proteomics literature, Limma and SAM perform better than Student's. However, Student t-test remains useful to represent baseline performances.

Limma and SAM lead to similar performances As both our method and the SAM test are based on the same regularization principle, we use SAM in our comparisons to represent the state-of-the-art performances. We note however that limma accepts more complex hierarchical designs while the native implementation of SAM cannot represent technical batches.

SAM is classically used in proteomics by using the fudge factor to mimic a threshold on the fold change and picking the value that leads to the best detection performance on a dataset, as discussed in Gianetto *and others* (2016). Such use is invalid as it amounts to overfitting and leads to over-optimistic results.

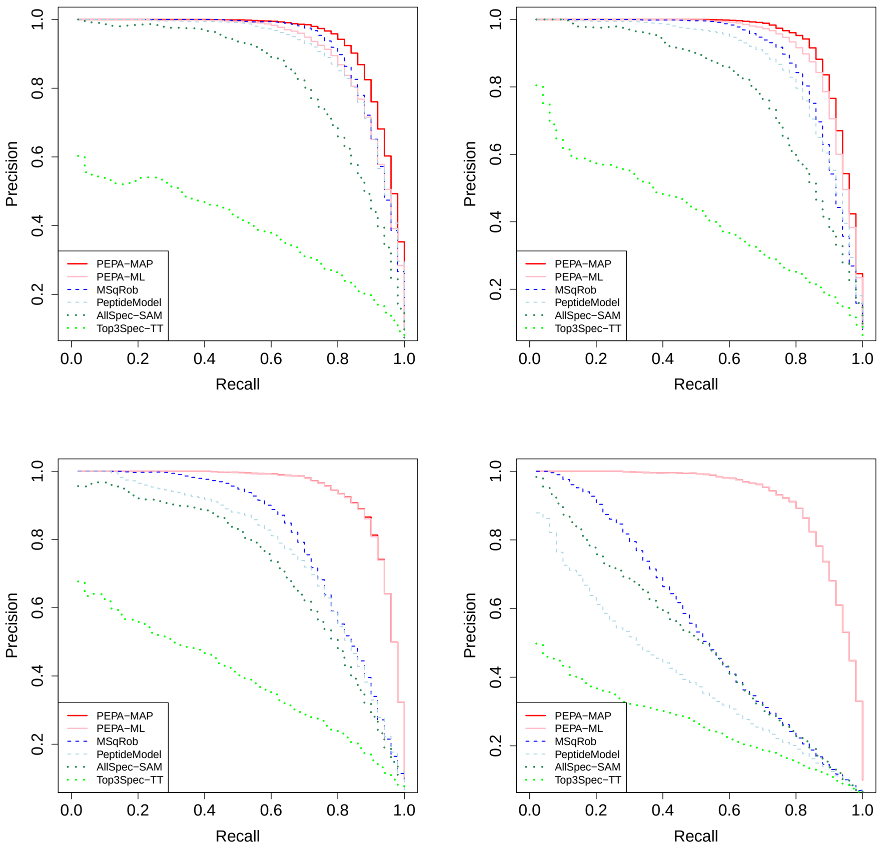


Figure 3: PR curve on simulated data with 1% (upper left), 10% (upper right), 33% (lower left) and 67% (lower right) of shared peptides.

Within this work, we only consider the automatic tuning of the fudge factor that is described in the original SAM publication Tusher *and others* (2001).

D Additional plots

See Figure 3 and the following ones.

References

Gerster, Sarah, Kwon, Taejoon, Ludwig, Christina, Matondo, Mariette, Vogel, Christine, Marcotte, Edward M, Aebersold, Ruedi

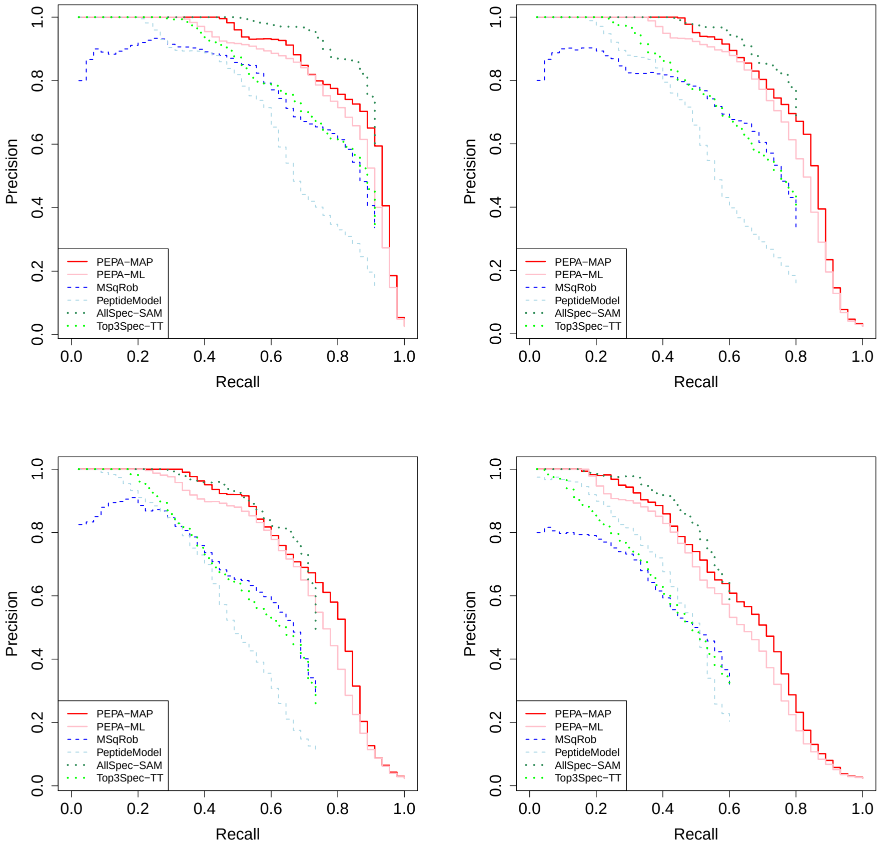


Figure 4: PR curve on Exp1_R2_pept data with 40 (upper left), 80 (upper right), 160 (lower left) and 240 (lower right) artificially added shared peptides.

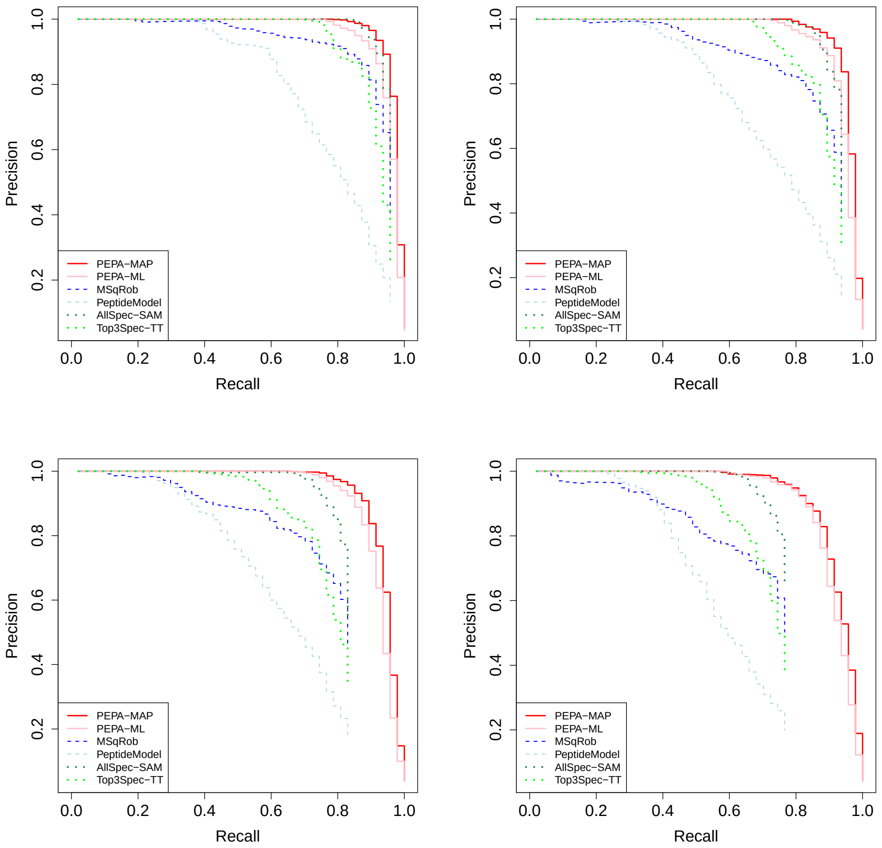


Figure 5: PR curve on Exp1_R25_pept data with 40 (upper left), 80 (upper right), 160 (lower left) and 240 (lower right) artificially added shared peptides.

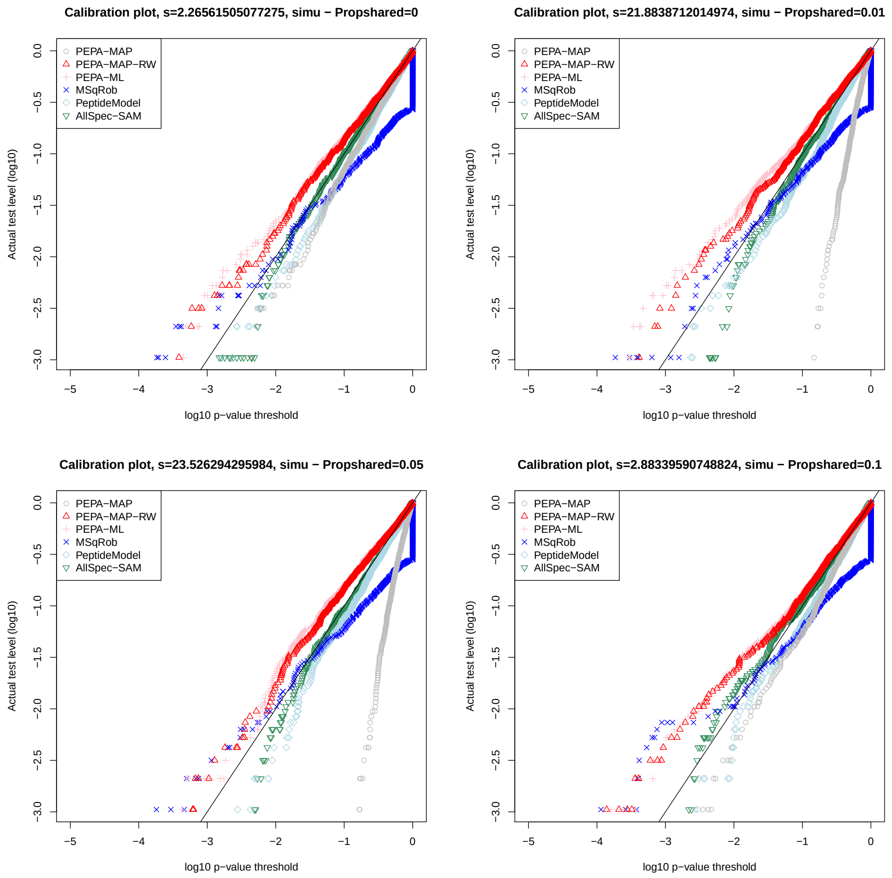
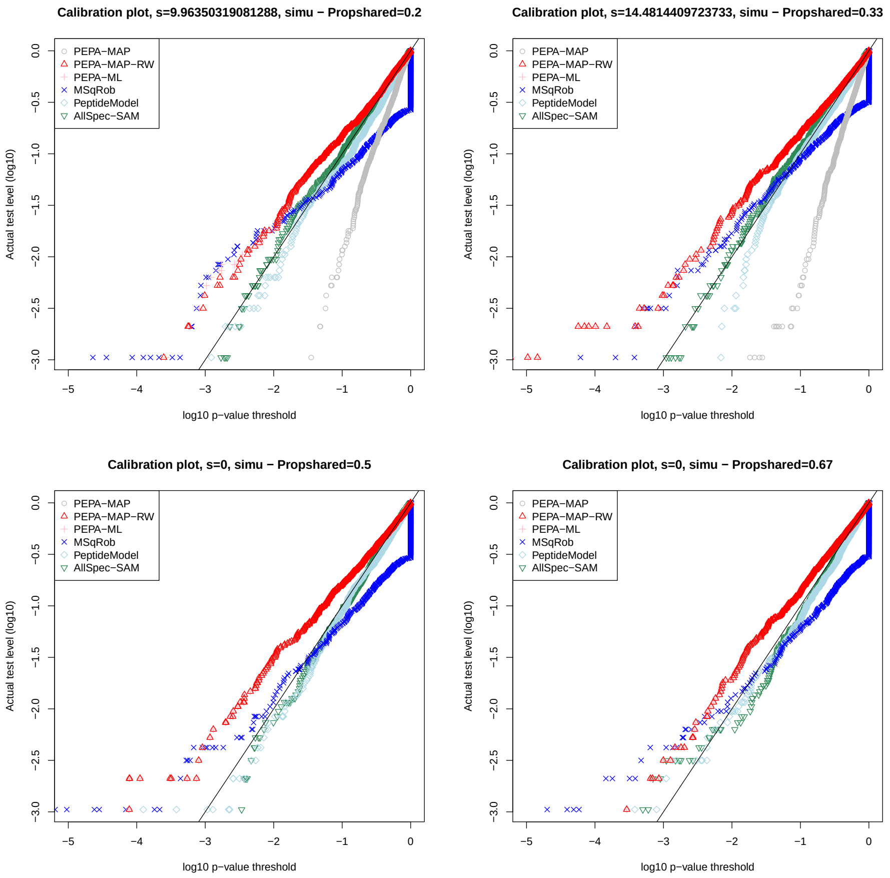


Figure 6: Calibration plots for simulated data (1/2).


Calibration plot, s=9.96350319081288, simu − Propshared=0.2

Calibration plot, s=14.4814409723733, simu − Propshared=0.33

Calibration plot, s=0, simu − Propshared=0.5

Calibration plot, s=0, simu − Propshared=0.67

Figure 7: Calibration plots for simulated data (2/2).

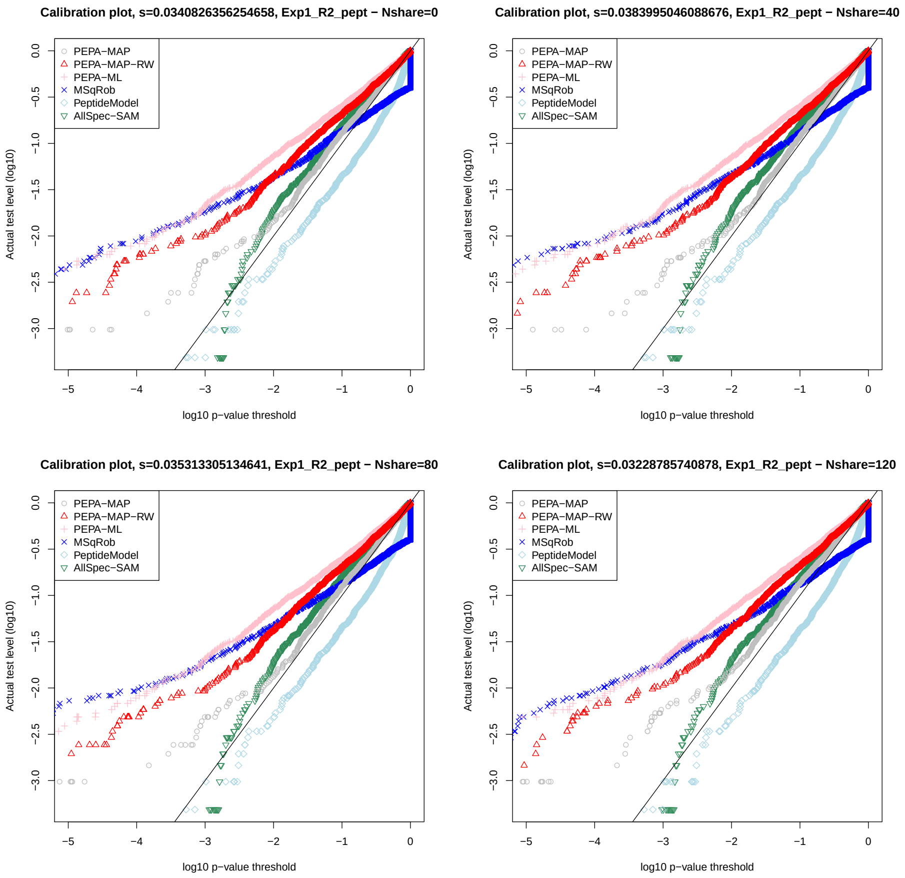


Figure 8: Calibration plots for Exp1_R2_pept data (1/2).

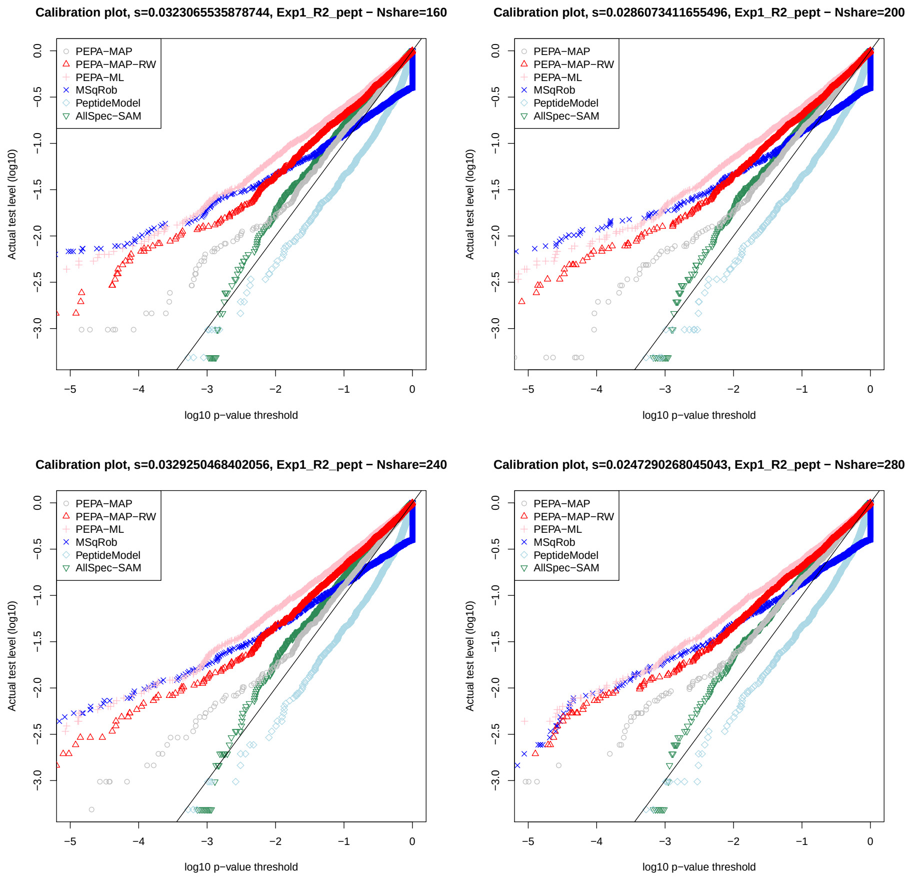
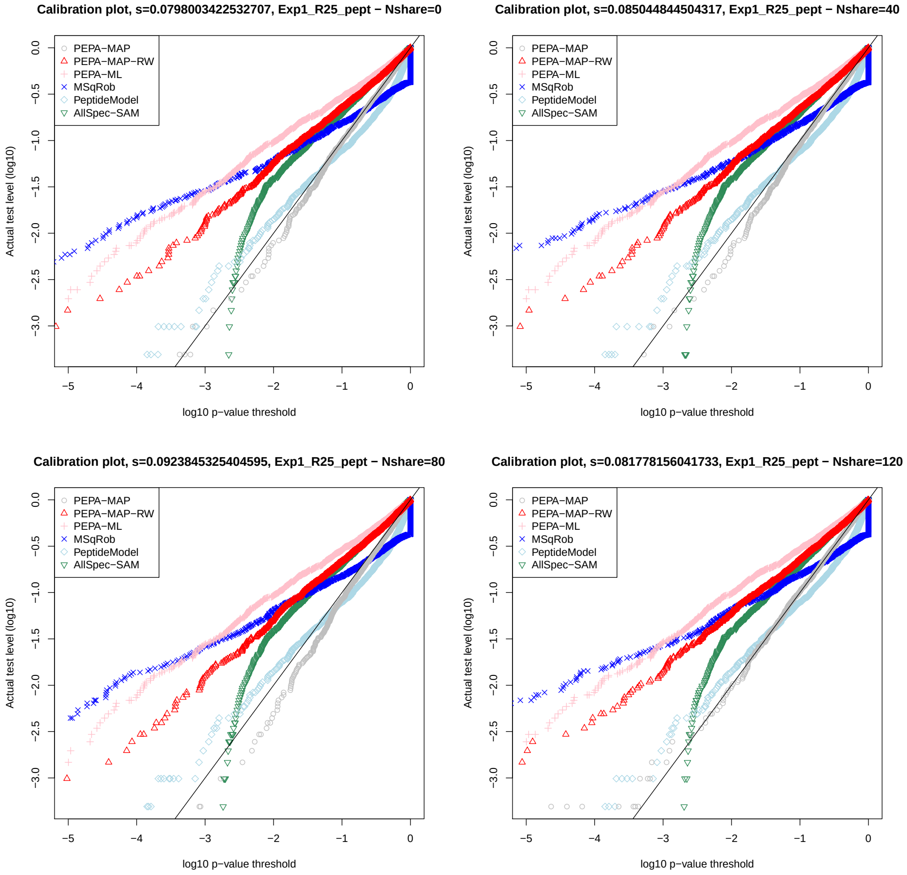


Figure 9: Calibration plots for Exp1_R2_pept data (2/2).

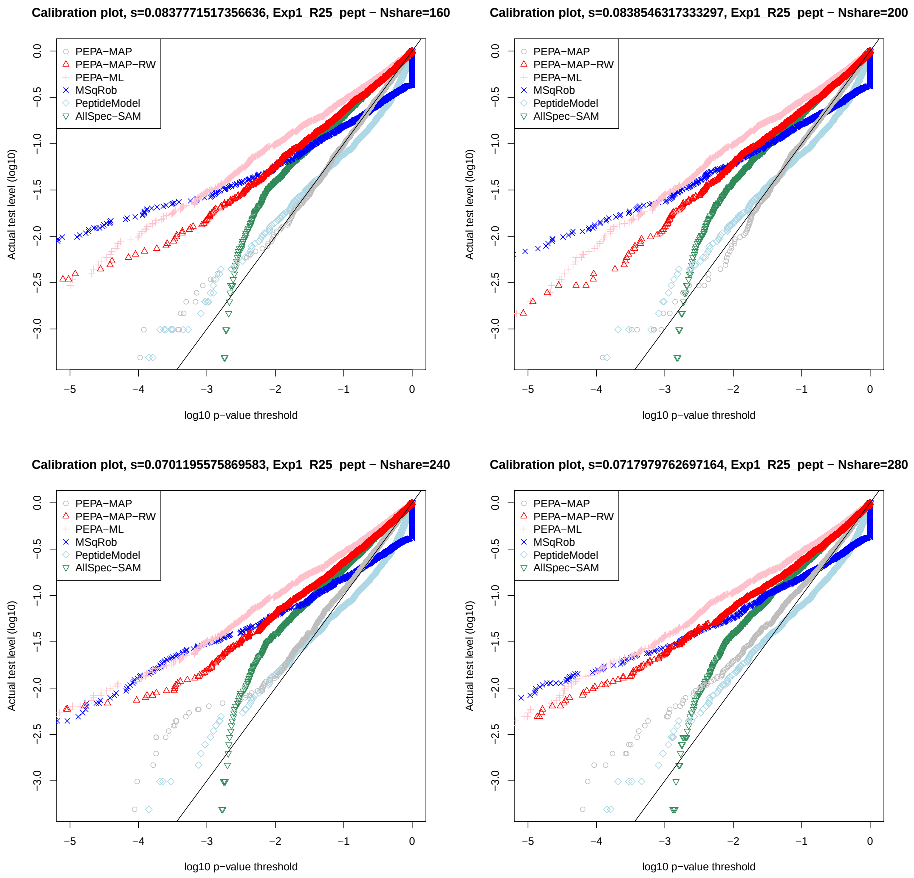

Calibration plot, s=0.0798003422532707, Exp1_R25_pept − Nshare=0

Calibration plot, s=0.085044844504317, Exp1_R25_pept − Nshare=40

Calibration plot, s=0.0923845325404595, Exp1_R25_pept − Nshare=80

Calibration plot, s=0.081778156041733, Exp1_R25_pept − Nshare=120

Figure 10: Calibration plots for Exp1_R25_pept data (1/2).



Calibration plot, s=0.0837771517356636, Exp1_R25_pept − Nshare=160

Calibration plot, s=0.0838546317333297, Exp1_R25_pept − Nshare=200

Calibration plot, s=0.0701195575869583, Exp1_R25_pept − Nshare=240

Calibration plot, s=0.0717979762697164, Exp1_R25_pept − Nshare=280

Figure 11: Calibration plots for Exp1_R25_pept data (2/2).

and Bühlmann, Peter. (2014). Statistical approach to protein quantification. *Molecular & cellular proteomics* **13**(2), 666–677.

Gianetto, Quentin Giai, Couté, Yohann, Bruley, Christophe and Burger, Thomas. (2016). Uses and misuses of the fudge factor in quantitative discovery proteomics. *Proteomics*.

Lazar, Cosmin, Gatto, Laurent, Ferro, Myriam, Bruley, Christophe and Burger, Thomas. (2016). Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *Journal of proteome research* **15**(4), 1116–1125.

Pham, Thang V, Piersma, Sander R, Warmoes, Marc and Jimenez, Connie R. (2010). On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. *Bioinformatics* **26**(3), 363–369.

Silva, Jeffrey C, Gorenstein, Marc V, Li, Guo-Zhong, Vissers, Johannes PC and Geromanos, Scott J. (2006). Absolute quantification of proteins by lcmse a virtue of parallel ms acquisition. *Molecular & Cellular Proteomics* **5**(1), 144–156.

Smyth, Gordon K. (2005). Limma: linear models for microarray data. In: *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, pp. 397–420.

Ting, Lily, Cowley, Mark J, Hoon, Seah Lay, Guilhaus, Michael, Raftery, Mark J and Cavicchioli, Ricardo. (2009). Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling. *Molecular & Cellular Proteomics* **8**(10), 2227–2242.

Tusher, Virginia Goss, Tibshirani, Robert and Chu, Gilbert. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98**(9), 5116–5121.