BIOINFORMATICS

Pages 1–2

Supplementary Materials: Detecting differential protein expression in large-scale population proteomics

So Young Ryu [1,3], Wei-Jun Qian [2], David G. Camp [2], Richard D. Smith [2], Ronald G. Tompkins [3], Ronald W. Davis [1] and Wenzhong Xiao [1,3,*]

[1]Stanford Genome Technology Center, Stanford University, Stanford, CA 94304, USA.
[2]Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA 99352, USA.
[3]Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA.

1 SIMULATION DETAILS

We simulated missing data using the following mechanisms:

$$
\begin{aligned}
g_{i1} &\sim \text{Uniform}(-5,5) \qquad (1)\\
g_{i1} &\sim \text{Uniform}(-5,-2)\\
u_i &\sim \text{Bernoulli}(b),\\
k_i &= \begin{cases} g_{i1} & \text{if } u_i = 1,\\ g_{i2} & \text{if } u_i = 0,\end{cases}\\
p_{i1} &= \Phi(-\tau_1(\mu_i - \bar{\mu}))\\
p_{i2} &= \Phi(\tau_2 k_i)\\
m_{i1} &\sim \text{Bernoulli}(p_{i1})\\
m_{i2} &\sim \text{Bernoulli}(p_{i2})\\
m_i &= \begin{cases} 0 & \text{if } m_{i1} = 0 \text{ and } m_{i2} = 0,\\ 1 & \text{otherwise.}\end{cases}
\end{aligned}
$$

Peptide i was absent when $m_i = 1$ and present when $m_i = 0$. Peptide i was absent when either its intensity was small ($m_{i1} = 1$) or the total number of quantified peptides was small compared to its abundance ranking ($m_{i2} = 1$). We allowed low abundant peptides to be present occasionally by generating m_{i1} from Bernoulli distribution. We generated m_{i2} in a similar fashion. These two missing mechanisms were determined by p_{i1} and p_{i2}. p_{i1} was large when a peptide intensity μ_i is small since $0 < \tau_1 \leq 1$. As mentioned in the main text, τ_1 was a magnitude of association between peptide intensities and missing rates and we varied this parameter to generate different scenarios. p_{i2} was generated based on k_i value. When k_i was generated from Uniform(-5,-2), it attempted to generate data from the part A in Figure 1 of the main paper. Otherwise, it attempted to generate data from the part B in Figure 1. b can correspond to the proportion of the part A and B in Fig 1 in the main text. A higher b value produced a larger portion of part B. We also varied the parameters, τ_2 and b, to generate various scenarios. In our simulated data, 2,500 proteins (out of 5,000) were different mean abundances between two groups.

2 MORE SIMULATION RESULTS

Performance Comparison. In Figure S1 and S2, we explored the performance of SALPS in the parameter space, (τ_1, τ_2, b). As mentioned in the main text, SALPS performed the best or close to the best. However, the performances of LinearI and LinearC varied depending on the simulation parameters. With small values of τ_1 and large b values, LinearC performed better than LinearI. With large τ_1 values and small b values, LinearI performed better than LinearC. With small τ_1 values, the performance of LinearI decreased as τ_2 increased.

Bootstrap vs. Likelihood Ratio Test. In SALPS, we estimated q-values using bootstrap approach coupled with Storey (2002). We can also use a likelihood ratio test instead of the bootstrap. However, q-values based on the bootstrap were closer to the actual q-values than the likelihood ratio test (Figure S3). Thus, we used bootstrap when testing $H_0 : \beta_g = \gamma_g = 0$.

The Signs of β_g and γ_g. After detecting differential proteins, we were interested in whether the differential proteins were more abundant in Group A or B. For each simulation, we generated 2,500 differential proteins with $\beta_g = -1$ and the other 2,500 proteins with no difference between two groups. Thus, the differential proteins correctly identified by the algorithm were expected to have negative β_g values, which implied that these proteins were more abundant in Group A, and positive γ_g values, which implied that these proteins had smaller missing rates in Group A. For most of differential proteins with their q-values< 0.01, the estimated $\widehat{\beta_g}$'s were negative and the estimated $\widehat{\gamma_g}$'s were positive (Figure S4a). Some differential proteins had incorrect signs in one of estimates, thus had $\widehat{\beta_g\gamma_g} > 0$.

For those proteins with $\widehat{\beta_g\gamma_g} > 0$, we further tested $H_{0\beta}$ and $H_{0\gamma}$ as proposed in the Method section and determined whether the differential proteins were more abundant in Group A or B. Figure S4b showed their test results. Based on these results, we determined whether proteins were more or less abundant in Group A than B based on $\widehat{\gamma_g}$ for blue circles and $\widehat{\beta_g}$ for green circles. (Only true positives were shown in Figure S4b to avoid confusion.) All blue circles located in Quadrant I with positive $\widehat{\gamma_g}$'s and all green circles located in Quadrant III with the negative $\widehat{\beta_g}$'s.

[*]to whom correspondence should be addressed.

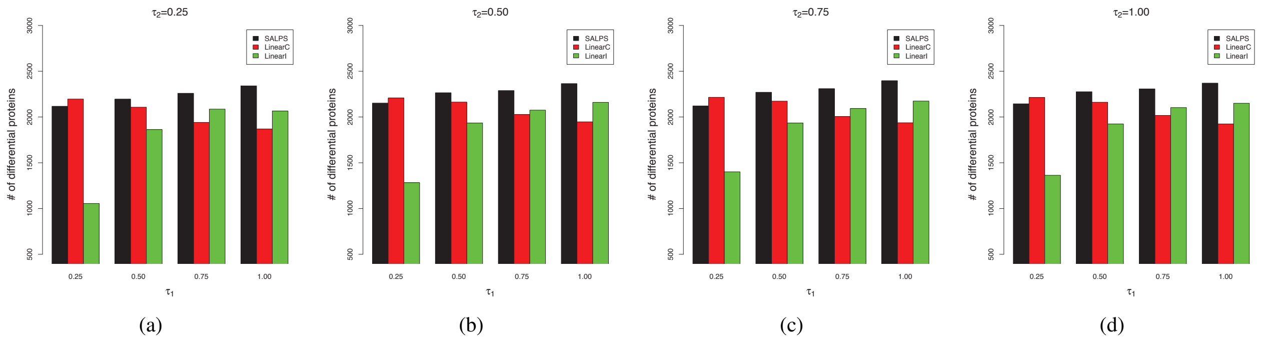


Fig. S1. Simulation results based on bootstrap approach and likelihood ratio test with various parameter values of τ_1 and τ_2 with a fixed value of $b = 0.50$. The bar height represented the numbers of differential proteins at $q < 0.01$.

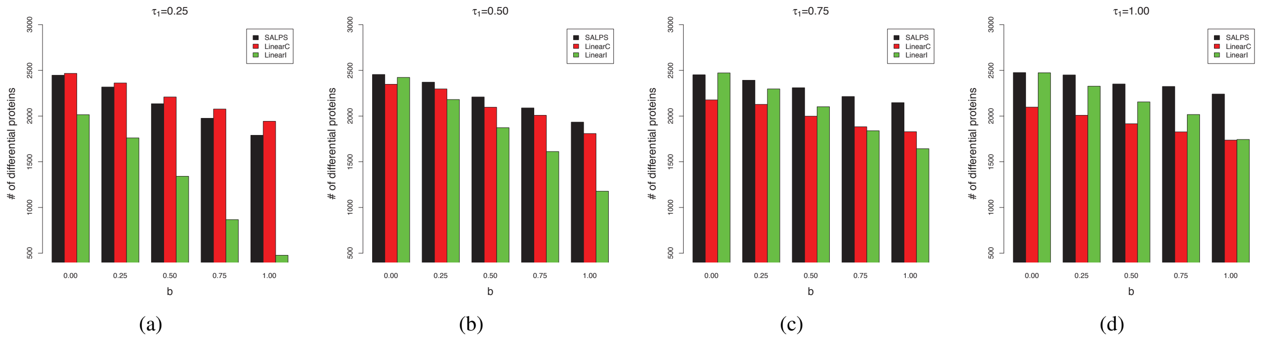


Fig. S2. Simulation results with various parameter values of b and τ_1 with a fixed value of $\tau_2 = 0.50$. The bar height represented the numbers of differential proteins at $q < 0.01$.

3 RESULTS USING ADDITIONAL REAL BIOLOGICAL DATA

Besides the monocytes proteomics data shown in the main text, we also compared SALPS with the alternative approaches using an additional monocytes proteomics data. Table S1 were the analysis results using monocytes proteomic samples collected one day after the injury. (Please note that monocytes proteomics samples used in the main text were collected 12 hours after the injury.) SALPS still performed better than the other two approaches, LinearC and LinearI, in terms of the number of differential proteins. However, this time, LinearC performed better than LinearI.

Table S1. The numbers of differential proteins between complicated vs. complicated patients one day after the injuries.

q-values	SALPS	LinearC	LinearI
$< 1e-04$	18	9	3
< 0.001	18	11	4
< 0.01	18	18	5
< 0.05	56	35	33

4 MORE COMPARISON RESULTS

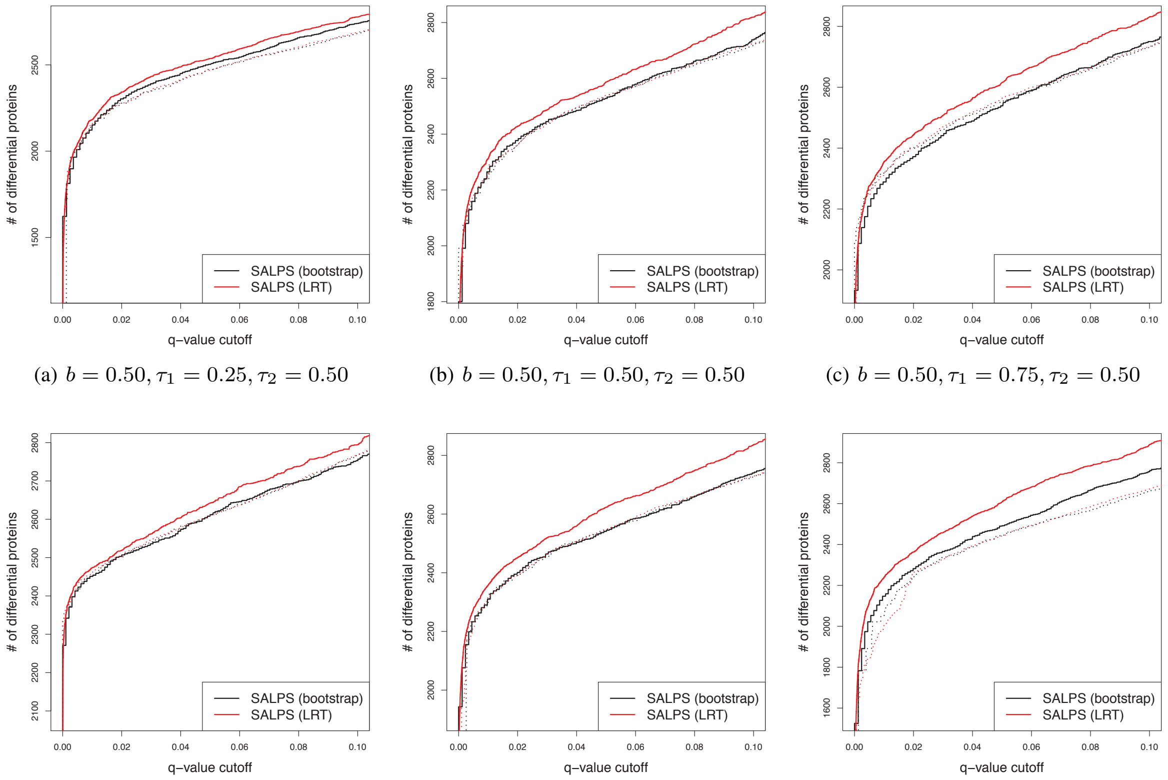
The comparison results of SALPS, LinearI, LinearC, Wang *et al.* (2012) and Karpievitch *et al.* (2009) using the simulated data sets are shown in Figure S5.

REFERENCES

Karpievitch, Y., Stanley, J., Taverner, T., Huang, J., Adkins, J. N., Ansong, C., Heffron, F., Metz, T. O., Qian, W. J., Yoon, H., Smith, R. D., and Dabney, A. R. (2009). A statistical framework for protein quantitation in bottom-up ms-based proteomics. *Bioinformatics*, **25**(16), 2028–34.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(3), 479–498.

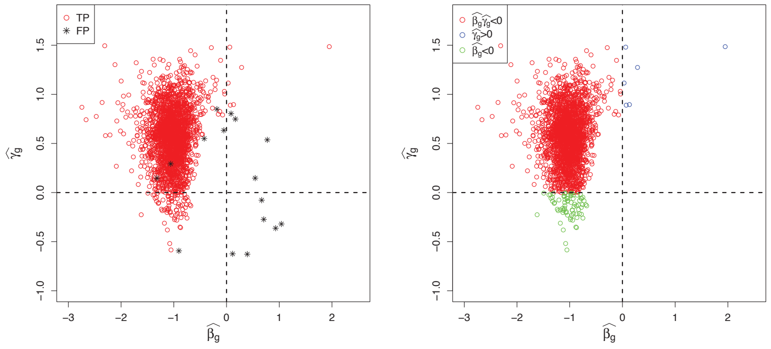
Wang, X., Anderson, G. A., Smith, R. D., and Dabney, A. R. (2012). A hybrid approach to protein differential expression in mass spectrometry-based proteomics. *Bioinformatics*, **28**(12), 1586–91.



(a) $b = 0.50, \tau_1 = 0.25, \tau_2 = 0.50$ (b) $b = 0.50, \tau_1 = 0.50, \tau_2 = 0.50$ (c) $b = 0.50, \tau_1 = 0.75, \tau_2 = 0.50$

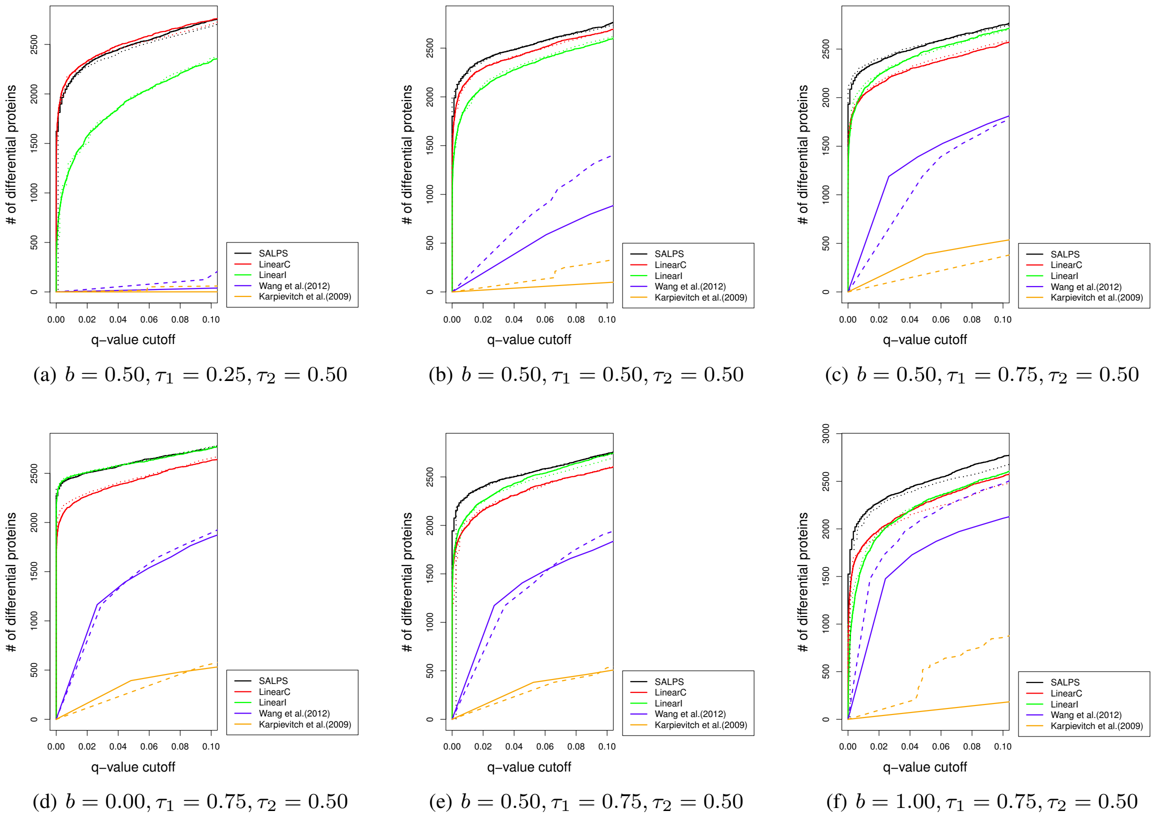
(d) $b = 0.00, \tau_1 = 0.75, \tau_2 = 0.50$ (e) $b = 0.50, \tau_1 = 0.75, \tau_2 = 0.50$ (f) $b = 1.00, \tau_1 = 0.75, \tau_2 = 0.50$

Fig. S3. Simulation results with various parameter values of b, τ_1, and τ_2. The solid lines were based on the estimated q-values while the dotted lines were based on the actual q-values.



(a) (b)

Fig. S4. Scatter plots between $\widehat{\beta_g}$'s and $\widehat{\gamma_g}$'s using the simulated data with $b = 0.50$, $\tau_1 = 0.75$, and $\tau_2 = 0.50$. (a) Red circles and black stars represented true (TP) and false (FP) differential proteins detected by SALPS at $q < 0.01$ respectively. (b) The true differential proteins detected by SALPS at $q < 0.01$. Red circles represented proteins with $\widehat{\beta_g} < 0$ and $\widehat{\gamma_g} > 0$. Green circles were proteins with their relative abundances determined by $\widehat{\beta_g}$. Blue circles were proteins with their relative abundances determined by $\widehat{\gamma_g}$.



(a) $b = 0.50, \tau_1 = 0.25, \tau_2 = 0.50$

(b) $b = 0.50, \tau_1 = 0.50, \tau_2 = 0.50$

(c) $b = 0.50, \tau_1 = 0.75, \tau_2 = 0.50$

(d) $b = 0.00, \tau_1 = 0.75, \tau_2 = 0.50$

(e) $b = 0.50, \tau_1 = 0.75, \tau_2 = 0.50$

(f) $b = 1.00, \tau_1 = 0.75, \tau_2 = 0.50$

Fig. S5. Simulation results showing the number of differentially expressed proteins vs. q-value at varying parameter values of (b, τ_1). b was the probabilities that missing values were generated from the total quantification-dependent missing mechanism. τ_1 was the magnitude of association between mean peptide intensities and missing rates. τ_2 was the magnitude of association between k values and missing rates ($\tau_2 = 0.50$). The solid lines were based on the estimated q-values while the dotted lines were based on the actual q-values.