Contents

1 Overview of proposed approach and notation

This study addresses three major goals in the characterization of post-translational modifications (PTMs): a) relative PTM quantification, b) PTM significance analysis, i.e., to detect PTM sites that are differentially modified across experimental conditions, and c) statistical design of PTM experiments.

Data structure of PTM quantification experiments. A set of fully-cleaved and/or partially-cleaved peptides containing a same PTM (e.g., ubiquitination) at one site are considered together. There are I conditions and J mass spectrometry runs (technical replicates) per condition in the experiment. The PTM site is represented by K spectral features (peptide ions, distinguished by their cleavage residues and charge states). The log-intensity (base 2) of Feature k, in Run j of Condition i is denoted by y_{ijk}. To account for the underlying protein abundance, features corresponding to the unmodified peptides from the same protein are considered together, except those unmodified peptides containing a modified site to avoid the confounding effect due to the PTM. The log-intensity of Feature l from the unmodified peptides in the same run is denoted by y^*_{ijl}. Figure S1 shows an example data representation of modified peptide ions at one site and unmodified peptide ions of the same protein. Unmodified peptides from the same protein provide additional evidence on the underlying protein abundance, which needs to be integrated for PTM characterization. To address the goals of PTM characterization, statistical analysis needs to summarize values in this table using appropriate statistical models, translate the goal into a model-based quantity of interest, and draw inference (i.e., characterize the uncertainty) about the quantity.

		Condition 1				...	Condition I			
		Run 1	Run 2	...	Run J	...	Run 1	Run 2	...	Run J
Modified	Feature 1	y_{111}	y_{121}	...	y_{1J1}	...	y_{I11}	y_{I21}	...	y_{IJ1}
	Feature 2	y_{112}	$-$	...	y_{1J2}	...	y_{I12}	y_{I22}	...	y_{IJ2}
	...	...	...	...	...	...	...	...	...	...
	Feature K	y_{11K}	y_{12K}	...	y_{1JK}	...	y_{I1K}	y_{I2K}	...	y_{IJK}
Unmodified	Feature 1	y^*_{111}	y^*_{121}	...	y^*_{1J1}	...	y^*_{I11}	y^*_{I21}	...	y^*_{IJ1}
	Feature 2	y^*_{112}	y^*_{122}	...	y^*_{1J2}	...	$-$	$-$	...	$-$
	...	...	...	...	...	...	...	...	...	...
	Feature L	y^*_{11L}	y^*_{12L}	...	y^*_{1JL}	...	y^*_{I1L}	y^*_{I2L}	...	y^*_{IJL}

Figure S1: Representation of the data of modified peptides at one site and unmodified peptides of the same protein, with I conditions and J replicate runs. Abundances of the PTM and protein are quantified by multiple spectral features (peptide ions, K for modified peptides and L for unmodified peptides). Some spectral features can be missing (shown as $-$), either randomly in individual runs or completely in certain conditions. In real practice, the number of runs can vary across conditions.

2 Existing method: two-sample t-test

Two-sample t-test is based on the null hypothesis that there is no difference in mean PTM abundance between Conditions i and i'. The abundance in each run is taken as input and is often estimated by sum of peak intensities. The t-test is typically performed based on the log of summarized value. For example, the log-abundance estimate for the PTM in Run j of Condition i is given by

$$\log\left(\sum_{k=1}^{K} 2^{y_{ijk}}\right).$$

For adjustment with respect to unmodified peptides, the estimate of PTM abundance is divided by the protein abundance estimate, and the t-test for the adjusted PTM abundance on log scale takes as input the difference of their log-estimates. The quantity is denoted by d_{ij} and is given by

$$d_{ij} = \log\left(\sum_{k=1}^{K} 2^{y_{ijk}}\right) - \log\left(\sum_{l=1}^{L} 2^{y^*_{ijl}}\right).$$

Alternatively, run-level summary to be described in Section 3.1.1 can also be used. The difference between the means of PTM abundance in Conditions i and i' are estimated as

$$\hat{\Delta} = \frac{1}{J} d_{i+} - \frac{1}{J} d_{i'+},$$

where $d_{i+} = \sum_{j=1}^{J} d_{ij}$ and the test statistic for the t-test is given by $\hat{\Delta}/\mathrm{SE}(\hat{\Delta})$. The statistical significance of the difference is determined by comparing the test statistic against the t distribution, with degrees of freedom $df = 2J - 2$ in balanced designs.

[TODO: If we use MSstats protein summarization + t-test, how much would they be different, methodologically and performance?]

3 Proposed approach

To characterize the observed feature intensities, different levels of variations are expressed using linear mixed models in consideration of the following factors: modification, condition, run, and feature. As different degrees of variability are present in the feature intensities of modified and unmodified peptides, they are expressed by separate models.

3.1 Statistical modeling and inference

The observed log-intensity of a modified peptide feature is denoted by y_{ijk} and represented as

$$y_{ijk} = \psi + C_i + R_{j(i)} + F_k + (R \times F)_{ijk},$$

where the effects of condition and feature are modeled as fixed effects:

$$\sum_{i=1}^{I} C_i = 0, \quad \sum_{k=1}^{K} F_k = 0,$$

and the effects of run and its interaction with feature are considered as random effects arising from normal distribution with mean 0:

$$R_{j(i)} = \gamma_{j(i)} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_\gamma), \qquad (R \times F)_{ijk} = \epsilon_{ijk} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_\epsilon).$$

Similarly, the observed log-intensity of an unmodified peptide feature is denoted by y^*_{ijl} and represented as

$$y^*_{ijl} = \psi^* + C^*_i + R^*_{j(i)} + F^*_l + (R \times F)^*_{ijl},$$

where the effects of condition and feature are modeled as fixed effects:

$$\sum_{i=1}^{I} C^*_i = 0, \quad \sum_{l=1}^{L} F^*_l = 0,$$

and

$$R^*_{j(i)} = \gamma^*_{j(i)} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\gamma^*}), \qquad (R \times F)^*_{ijl} = \epsilon^*_{ijl} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\epsilon^*}).$$

3.1.1 Run-level summarization of feature intensities

Run-level summarization of feature intensities for each PTM site is carried out as in the sub-plot model of MSstats (1), which involves a) imputation of censored missing values, and b) summarization of feature intensities using Tukey's median polish. The run-level summary for the PTM in Run j of Condition i is denoted by $\hat{y}_{ij}$.

3.1.2 Model-based inference of the underlying abundance

The PTM abundance in each run is represented as

$$\hat{y}_{ij} = \psi + C_i + R_{j(i)},$$

where $\sum_{i=1}^{I} C_i = 0$, $R_{j(i)} = \gamma_{j(i)} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_\gamma)$. Similarly, the protein abundance in each run is expressed as

$$\hat{y}_{ij} = \psi^* + C^*_i + R^*_{j(i)},$$

where $\sum_{i=1}^{I} C^*_i = 0$, $R^*_{j(i)} = \gamma^*_{j(i)} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\gamma^*})$. The expected values of log-abundances of the PTM and protein in Condition i are denoted by μ_i and μ^*_i, respectively, and the values are estimated as:

$$\hat{\mu}_i = \hat{\psi} + \hat{C}_i = \frac{1}{J}\hat{y}_{i+}$$
$$\hat{\mu}^*_i = \hat{\psi}^* + \hat{C}^*_i = \frac{1}{J}\hat{y}^*_{i+},$$

where the standard errors of the estimates are $\text{SE}(\hat{\mu}_i) = (\hat{\sigma}^2_\gamma / J)^{1/2}$ and $\text{SE}(\hat{\mu}^*_i) = (\hat{\sigma}^2_{\gamma^*}/J)^{1/2}$. Based on the estimates $\hat{\mu}_i$ and $\hat{\mu}^*_i$, the adjusted log-abundance of the PTM is given by $(\hat{\mu}_i - \hat{\mu}^*_i)$ and the standard error of the estimate is

$$\left[\frac{1}{J}\left(\hat{\sigma}^2_\gamma + \hat{\sigma}^2_{\gamma^*}\right)\right]^{1/2}.$$

3.2 PTM significance analysis

With protein-level adjustment, the model-based testing is based on the hypothesis that there is no difference in adjusted PTM abundance between Conditions i and i'

$$H_0 : \Delta = (\mu_i - \mu_{i'}) - (\mu_i^* - \mu_{i'}^*) = 0$$
$$H_a : \Delta = (\mu_i - \mu_{i'}) - (\mu_i^* - \mu_{i'}^*) \neq 0$$

The log-fold change in the adjusted PTM abundance, Δ, is estimated by

$$\hat{\Delta} = \left[\frac{1}{J}\left(\hat{y}_{i+} - \hat{y}_{i'+}\right)\right] - \left[\frac{1}{J}\left(\hat{y}_{i+}^* - \hat{y}_{i'+}^*\right)\right],$$

and the standard error of the estimate $\mathrm{SE}(\hat{\Delta})$ is

$$\left[\frac{2}{J}\left(\hat{\sigma}_{\gamma}^2 + \hat{\sigma}_{\gamma^*}^2\right)\right]^{1/2}.$$

The test statistic $\hat{\Delta}/\mathrm{SE}(\hat{\Delta})$ is compared against the t distribution, with degrees of freedom approximated by

$$\left(\hat{\sigma}_{\gamma}^2 + \hat{\sigma}_{\gamma^*}^2\right)^2 \Big/ \left(\frac{\hat{\sigma}_{\gamma}^4}{\mathrm{df}(\gamma)} + \frac{\hat{\sigma}_{\gamma^*}^4}{\mathrm{df}(\gamma^*)}\right).$$

A distinctive property of the proposed model-based testing to the two-sample t-test is that even only the PTM abundances in Conditions i and i' are compared, measurements from all conditions are used for the modeling and inference.

3.3 Design of PTM experiments

The proposed statistical framework allows for design of PTM experiments in terms of sample size calculation and power analysis. Sample size calculation takes as input a) q, the desired false discovery rate, b) β, the average Type II error rate, c) Δ, the minimal log-fold change in adjusted PTM abundance that we would like to detect, d) $m_0/(m_0+m_1)$, the fraction of truly differentially modified PTM sites in the comparison, and e) σ_{γ}^2 and $\sigma_{\gamma^*}^2$, the anticipated variances associated to modified and unmodified peptide features, respectively. The variances can be derived based on the dataset being analyzed, assuming similar quantitative properties and variations. With these values and a user-specified number of conditions, the corresponding number of technical replicates per condition can then be derived, as described in (2). Given the above quantities, the minimal number of replicates J is determined by the variance of the estimated log-fold change $\mathrm{SE}^2(\hat{\Delta})$ as

$$\mathrm{SE}^2(\hat{\Delta}) = \left[\frac{2}{J}\left(\hat{\sigma}_{\gamma}^2 + \hat{\sigma}_{\gamma^*}^2\right)\right] \leq \left(\frac{\Delta}{t_{1-\beta,df} + t_{1-\alpha/2,df}}\right)^2,$$

where

$$\alpha = (1-\beta) \cdot \frac{q}{1 + (1-q) \cdot m_0/m_1},$$

and $t_{1-\beta,df}$ and $t_{1-\alpha/2,df}$ are the $100(1-\beta)^{\text{th}}$ and the $100(1-\alpha/2)^{\text{th}}$ percentiles of the t distribution, with $df = I(J-1)$ degrees of freedom in balanced designs. More details can be found in (3).

3.4 Extension to complex designs

The proposed statistical framework can be extended to analyze data from experiments of complex designs, such as factorial design and time series. We discuss below a specific design commonly applied in PTM experiments, in which data are acquired in multiple batches.

3.4.1 Batch effects

As in Section 3.2, hypothesis testing on the adjusted log-abundances of the PTM in Conditions i and i' is performed to detect differentially modified PTM sites. The hypothesis is that there is no difference in adjusted log-abundance of the PTM between Conditions i and i'

$$H_0 : \Delta = (\mu_i - \mu_{i'}) - (\mu_i^* - \mu_{i'}^*) = 0$$
$$H_a : \Delta = (\mu_i - \mu_{i'}) - (\mu_i^* - \mu_{i'}^*) \neq 0$$

For batch-wise data, we consider two ways to estimate the difference in adjusted PTM abundance, $\Delta = (\mu_i - \mu_{i'}) - (\mu_i^* - \mu_{i'}^*)$ based on different assumptions about the properties of batch effects, namely per-batch model (proposed approach) and all-batch model. For the following discussion, we denote the log-intensity of a modified peptide feature in Run j of Condition i and Batch b by $y_{b,ijk}$, where $b = 1, \ldots, B$. Similarly, $y_{b,ijl}^*$ is denoted for the unmodified peptide feature in the same run.

Per-batch model (proposed approach). The model assumes different levels of variability are present in different batches and the differences between conditions vary across batches (i.e., there is an interaction effect between condition and batch). Difference between conditions is estimated in each batch, and the overall log-fold change in adjusted PTM abundance is estimated as the average over batches

$$\hat{\Delta} = \frac{1}{B}\sum_{b=1}^{B}\left[\frac{1}{J}\left(\hat{y}_{b,i+} - \hat{y}_{b,i'+}\right)\right] - \frac{1}{B}\sum_{b=1}^{B}\left[\frac{1}{J}\left(\hat{y}_{b,i+}^* - \hat{y}_{b,i'+}^*\right)\right].$$

The standard error associated to the estimate is

$$\left[\left(\frac{1}{B}\right)^2 \cdot \left(\frac{2}{J}\right) \cdot \sum_{b=1}^{B}\left(\hat{\sigma}_{\gamma_b}^2 + \hat{\sigma}_{\gamma_b^*}^2\right)\right]^{1/2}.$$

The test statistic $\hat{\Delta}/\mathrm{SE}(\hat{\Delta})$ is compared against the t distribution, where the degrees of freedom are approximated as

$$\left[\sum_{b=1}^{B}\left(\hat{\sigma}_{\gamma_b}^2 + \hat{\sigma}_{\gamma_b^*}^2\right)\right]^2 \Big/ \sum_{b=1}^{B}\left[\frac{\hat{\sigma}_{\gamma_b}^4}{\mathrm{df}(\gamma_b)} + \frac{\hat{\sigma}_{\gamma_b^*}^4}{\mathrm{df}(\gamma_b^*)}\right].$$

All-batch model. In contrast to the per-batch model, the all-batch model assumes identical variance and difference between conditions in all batches. The log-fold change in adjusted PTM abundance is estimated as the average over runs and batches

$$\hat{\Delta} = \frac{1}{BJ}\left(\hat{y}_{+,i+} - \hat{y}_{+,i'+}\right) - \frac{1}{BJ}\left(\hat{y}_{+,i+}^* - \hat{y}_{+,i'+}^*\right),$$

and the standard error of the estimate is

$$\left[\frac{2}{BJ}\left(\hat{\sigma}_{\gamma}^{2}+\hat{\sigma}_{\gamma^{*}}^{2}\right)\right]^{1/2}.$$

The test statistic $\hat{\Delta}/\mathrm{SE}(\hat{\Delta})$ is compared against the t distribution, with degrees of freedom approximated by

$$\left(\hat{\sigma}_{\gamma}^{2}+\hat{\sigma}_{\gamma^{*}}^{2}\right)^{2} \Bigg/ \left(\frac{\hat{\sigma}_{\gamma}^{4}}{\mathrm{df}(\gamma)}+\frac{\hat{\sigma}_{\gamma^{*}}^{4}}{\mathrm{df}(\gamma^{*})}\right).$$

3.4.2 TMT experiment

[TODO: add the method for TMT experiment, batch vs plex?]

4 Computer simulation

The proposed statistical approach was evaluated and compared to the t-test using computer simulation. In particular, their statistical properties with respect to protein-level adjustment and batch effects were evaluated.

4.1 Protein-level adjustment

4.1.1 One site per protein

Differential levels of modified peptides may be due to differential modifications and/or changes in protein abundance. The proposed approach adjusts the PTM abundance with respect to protein abundance as introduced in Section 3. Alternatively, two-sample t-test taking as input the ratio between feature intensities of modified and unmodified peptides (difference on log scale) is commonly applied for the same purpose (Section 2). Approaches without considering protein-level adjustment lose track of an important aspect in interpreting observed changes in PTM abundance, which may result in misleading conclusions. To highlight the necessity of the adjustment, we compared the following approaches: a) proposed approach, b) linear model (no adjustment), i.e., same as the proposed approach but ignoring the measurements with respect to unmodified peptides, c) t-test (with adjustment), and d) t-test (no adjustment). In experiments of complex designs, multiple inter-related conditions are often compared together. Whereas t-test uses measurements from the two conditions being compared, the proposed approach leverages measurements in all conditions for the inference of the underlying abundance. To highlight this distinction, multiple conditions of data were generated, where only one condition was simulated with a systematic change. The simulation was based on the following parameters:

- Mean of log-intensity: 25
- Standard deviations of log-intensities for modified and unmodified peptides: 0.2, 0.3
- Difference in PTM abundance between conditions: 0, 0.5, 0.75, 1
- Difference in protein abundance between conditions: 0, 0.5
- Number of replicates: 2, 3, 5
- Number of conditions: 2, 3, 4
- Number of realizations: 500
- Missing data: no missing value, or PTM missing in Run 1, Condition 1

The results are summarized from Figure S2 to Figure S4, including false positive rate with or without changes in protein abundance by the considered methods (Figure S2), power by the approaches with protein-level adjustment when there is no missing data (Figure S3), and power by the approaches with protein-level adjustment when the PTM is missing in one run (Figure S4).

4.1.2 Multiple sites per protein

Expression levels of PTM sites are adjusted based on the abundance of their originating protein. Since the same reference is used for all sites, it introduces correlation among estimates and test statistics for those sites. This may cause issues in controlling false discovery rate (FDR). We investigated the property by simulating data of two conditions and 1000 proteins with different numbers of PTM sites and comparing the results of the proposed approach and the linear model with no adjustment. A fraction (50%) of the 1000 proteins had no changes between conditions while systematic changes were simulated for the rest of the proteins. Multiple testing correction was performed using the Benjamini and Hochberg's method. Performances of the considered approaches were assessed by their actual FDR, calculated as the fraction of proteins with adjusted p-values < 0.05 among the proteins with true differences. The results are summarized in Figure S5.

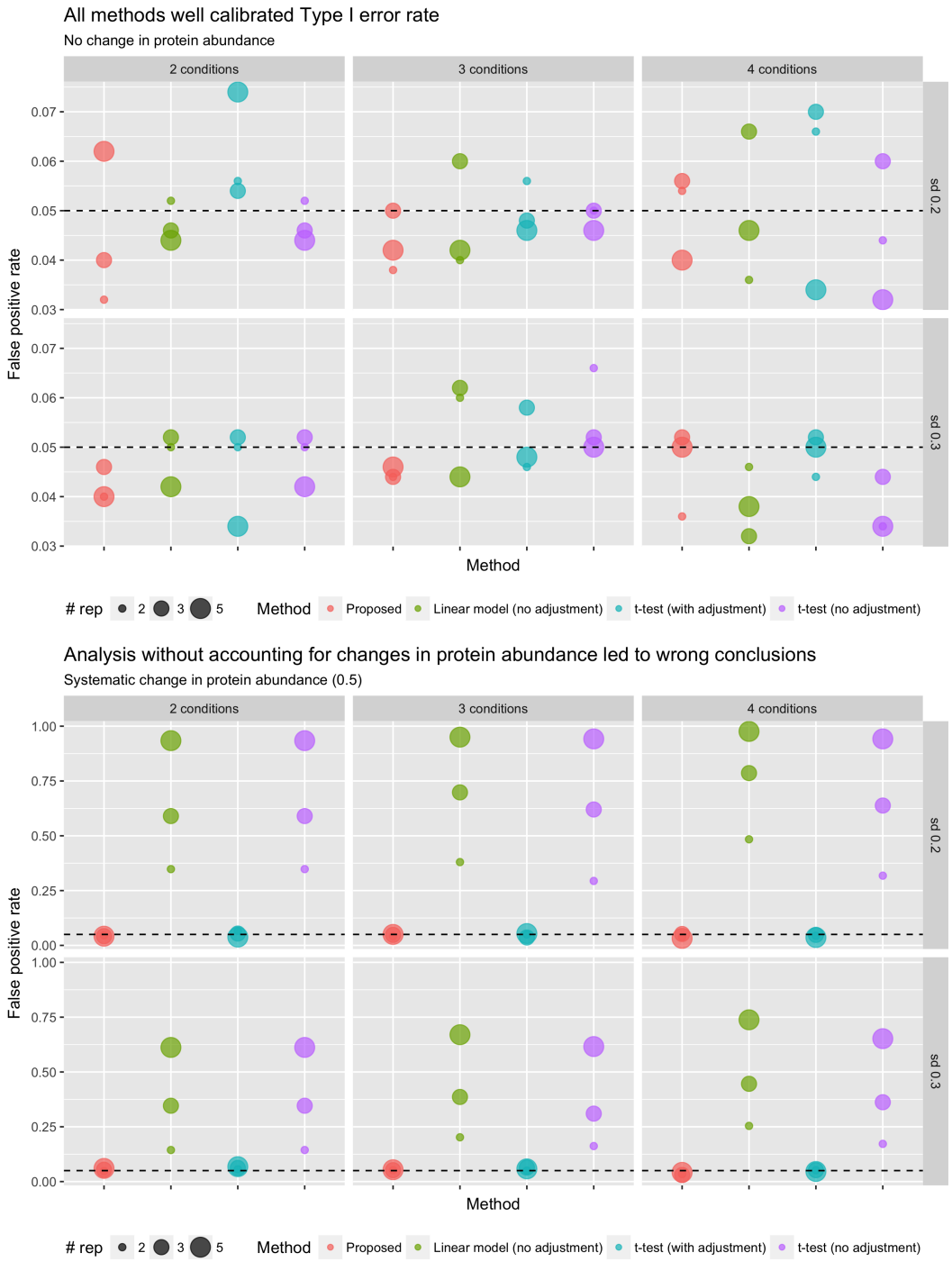

Figure S2: All the considered methods well calibrated the Type I error rate when there was no change in protein abundance (upper plot). When the changes in PTM abundance were entirely due to changes in protein abundance across conditions (bottom plot), analysis without accounting for the protein-level changes resulted in off-target, high false positive rates.

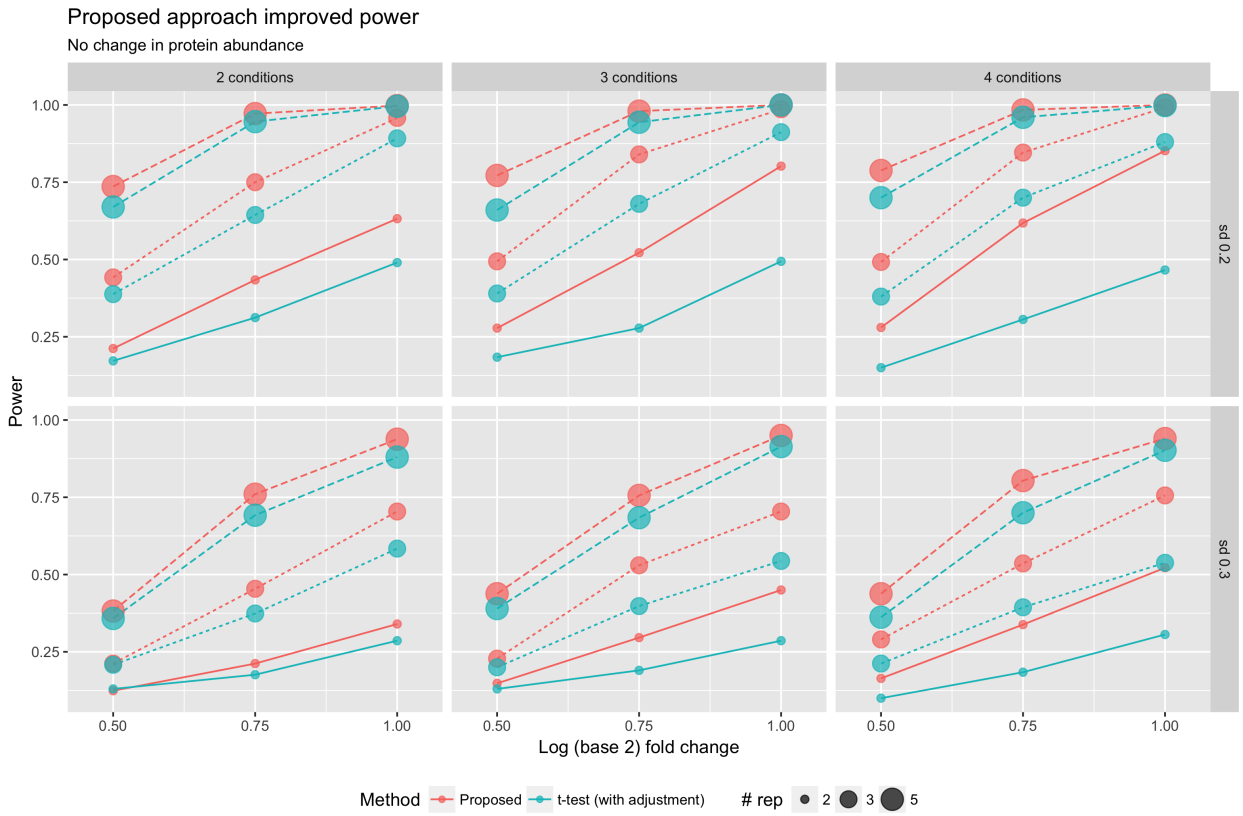

Figure S3: In comparison between the proposed approach and t-test with protein-level adjustment, the proposed approach improved power with small sample sizes. Two-sample t-test only used data within the groups of interest while ignoring the rest the of data. Consequently, it gave similar performance across cases with different number of conditions. In contrast, the proposed approach leveraged all available information, which resulted in improved power with increased number of conditions.

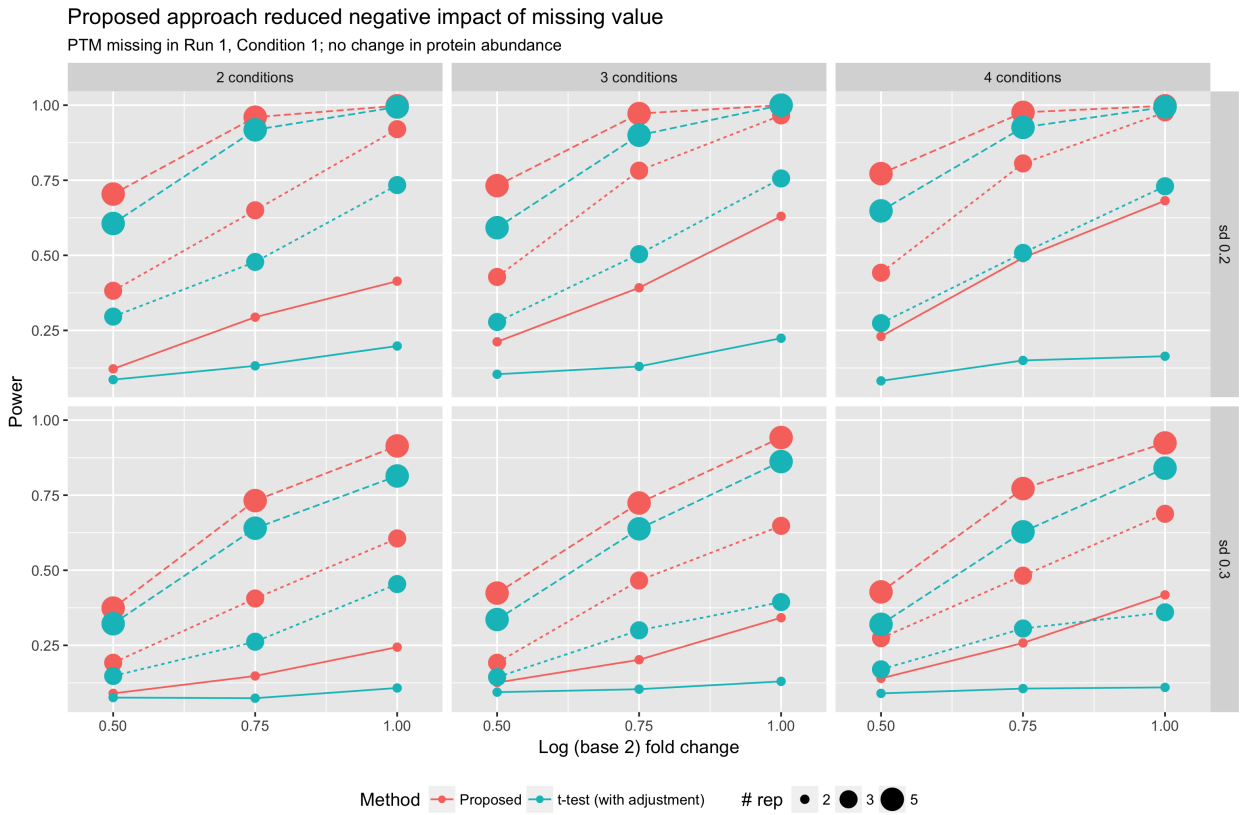

Figure S4: The advantage by using the proposed approach over two-sample t-test became more profound in the presence of missing value. In cases with small sample sizes, performance by the t-test decreased dramatically with one missing value. As a specific example, the statistical power by using the t-test with two replicates was below 0.25, and increasing the fold change to 2 did not effectively reduce the negative impact.

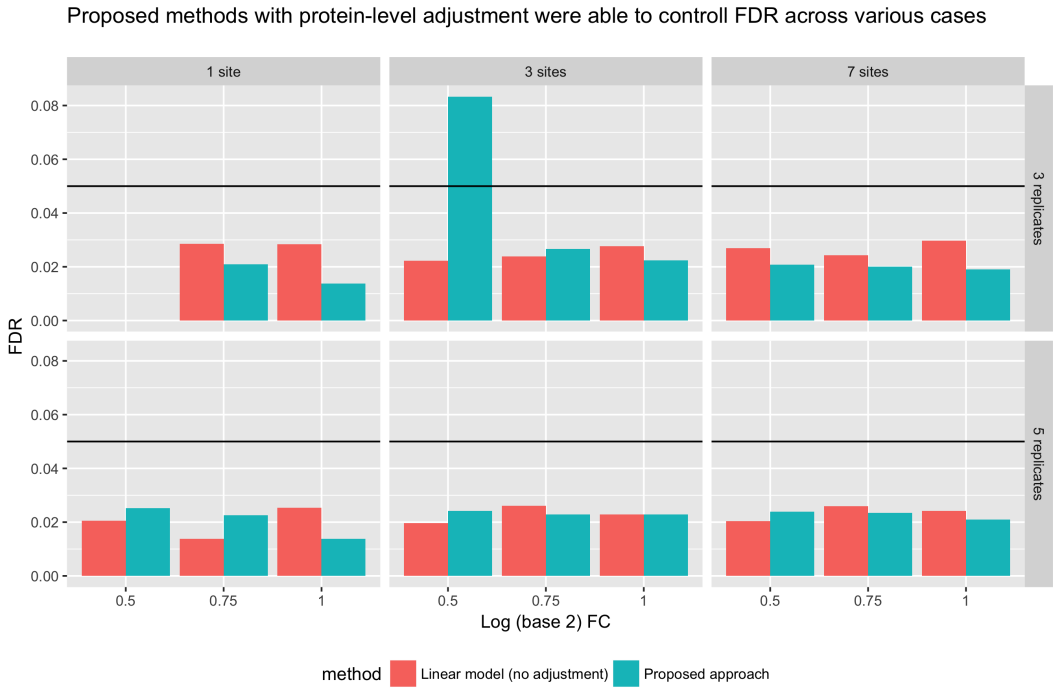

Figure S5: Correlation between test statistics across sites due to protein-level adjustment did not affect the expected level of FDR under the considered scenarios. The Benjamini-Hochberg procedure was used to control the FDR.[TODO: make same color for each method]

4.2 Batch effects

We evaluated the statistical properties of the proposed approach under batch effects, in comparison to the two-sample t-test (Section 2). While the t-test is not directly applicable to a problem with batches of data, several *ad-hoc* approaches may be used. Two commonly used approaches are a) t-test (no batch): ignoring batch effects when applying t-test, and b) t-test (most significant batch): applying t-test in each batch and drawing conclusions based on the most significant batch. Although simple, these *ad-hoc* methods lack statistical justification. We characterized their statistical properties under various forms of batch effects. In this part of the simulation, two batches of data were generated, with the following forms of batch effects: difference in signal intensities across batches, difference in variability across batches, and interaction effect between batch and condition (i.e., change between condition affected by batch). The following interaction scenarios were considered in the simulation: a) no interaction between condition and batch, i.e., same change across conditions in both batches, b) positive interaction as 25% greater change in the batch of higher level, and c) negative interaction as 25% lower change in the batch of higher level. Below are the parameters used in the simulation:

- Mean of log-intensity: 25
- Increased intensity level of Batch 2 versus Batch 1: 0, 1, 2
- Standard deviations of log-intensity: 0.2 in Batch 1 and 0.3 in Batch 2
- Difference between conditions in Batch 1: 0.5, 0.75, 1
- Number of replicates: 2, 3, 5
- Number of conditions: 2, 3, 4
- Number of realizations: 500

The following four approaches were compared: a) proposed approach, i.e., per-batch model, b) all-batch model, c) t-test (no batch), and d) t-test (most significant batch). The results are summarized from Figure S6 to Figure S9, on the aspects of estimation error (Figure S6), false positive rate (Figure S7), and power (Figure S8, Figure S9).

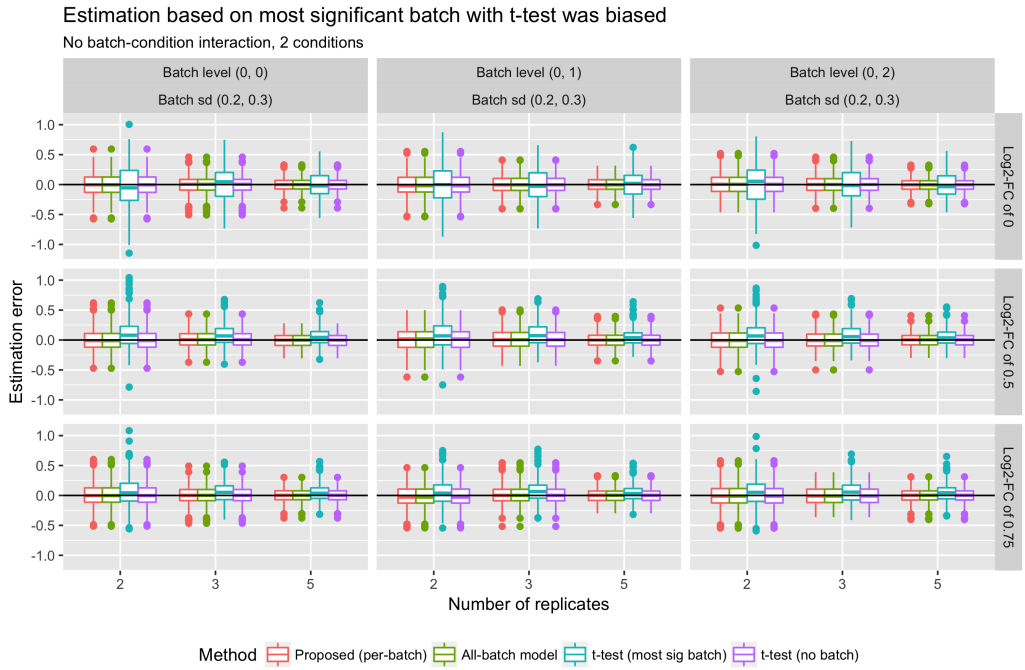

Figure S6: Estiamtion based on the most statistically significant batch with t-test was highly variable and frequently biased. The observation is consistent in all the simulated scenarios.

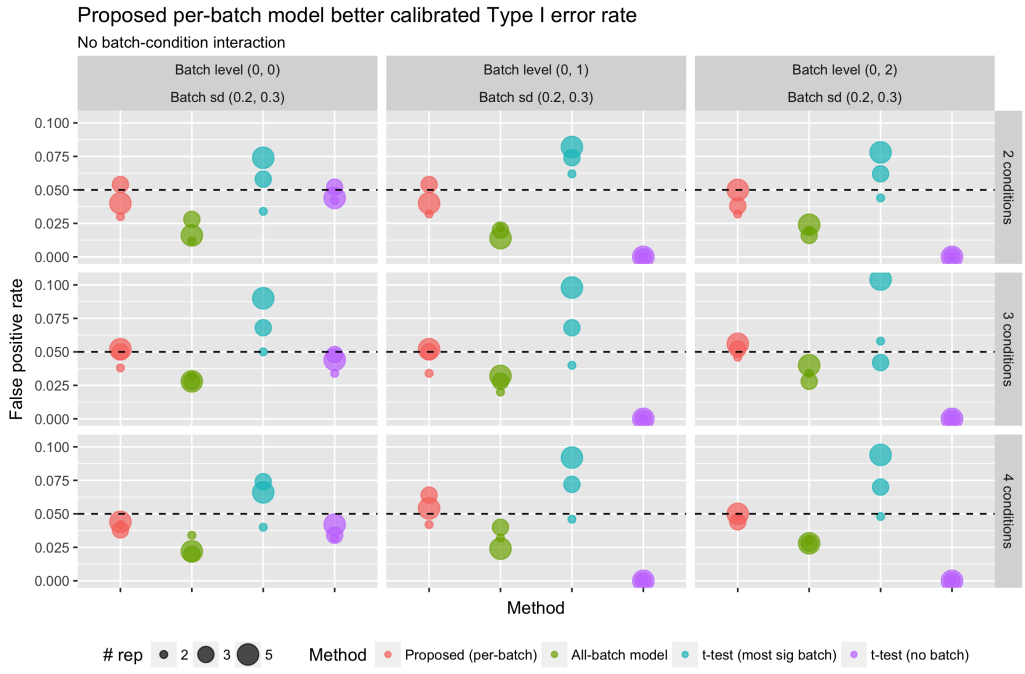

Figure S7: The proposed approach better calibrated Type I error rate. Similar results were observed in the cases with positive and negative batch-condition interactions.

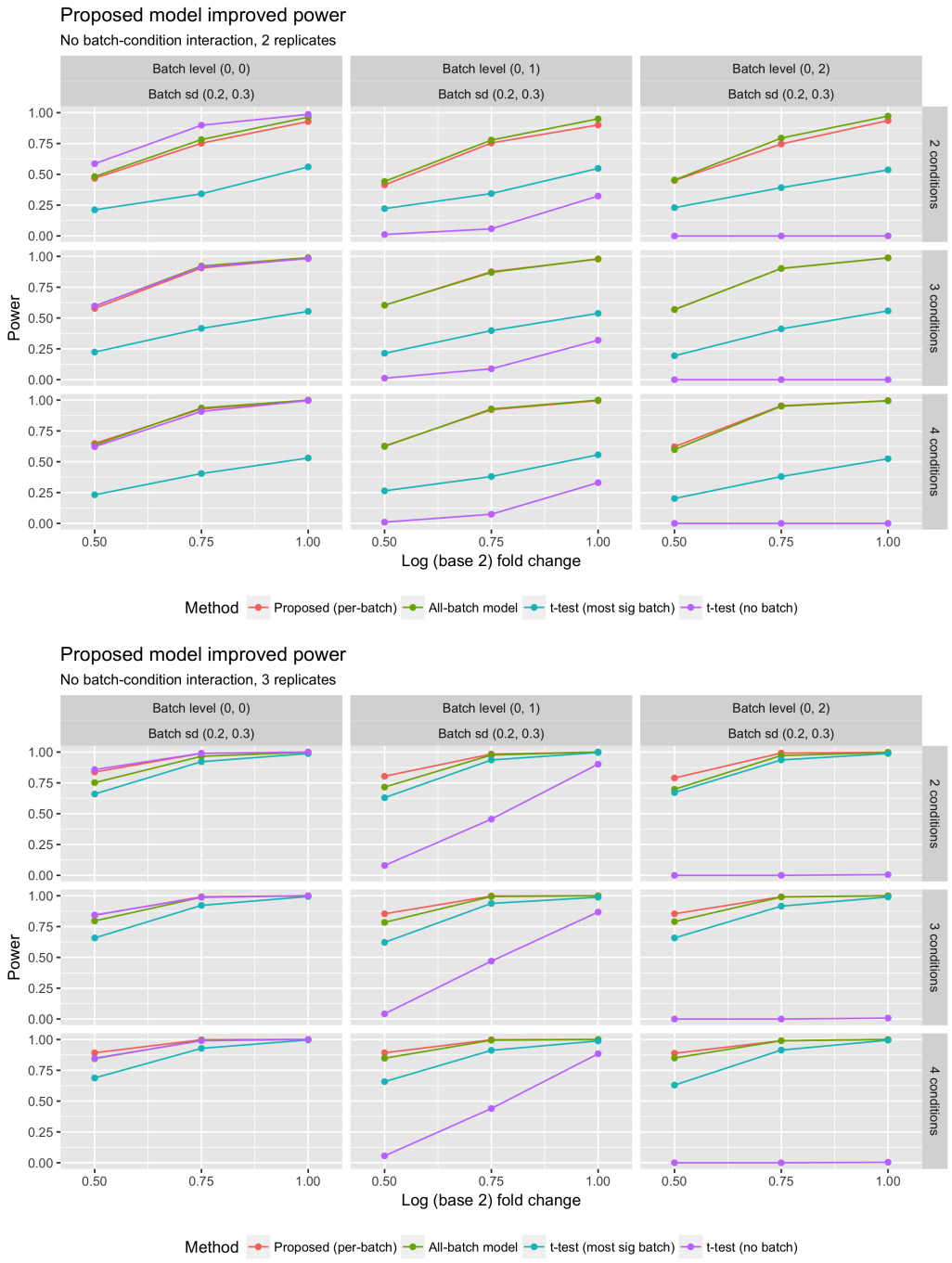

Figure S8: The proposed approach improved power with small sample sizes in almost all the scenarios, under various forms of batch effects including difference in intensity level (higher on the right) and difference in variability. Using t-test while ignoring batch effect gave similar performance in special cases with no difference in intensity level between batches, but its performance dramatically decreased in general cases. The proposed methods gave consistently improved performance compared with other methods by properly characterizing batch effects and leveraging all available information. Similar patterns were observed in other simulated scenarios with positive and negative interactions.

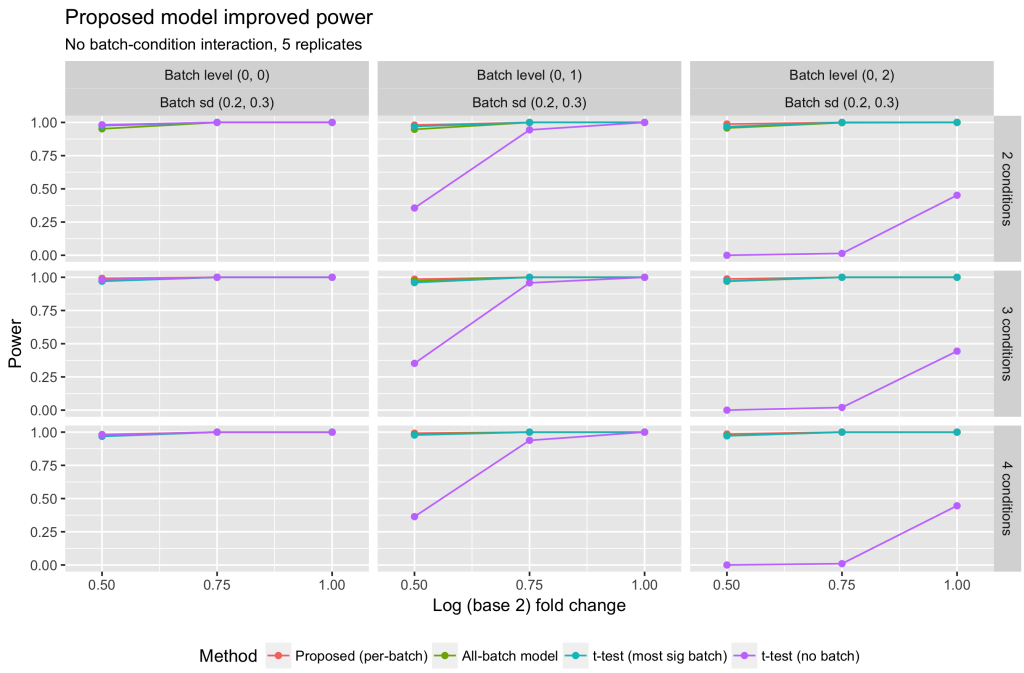

Figure S9: Ignoring batch effects lost power. The negative impact was only partially reduced by increasing the sample size to 5. Similar patterns were observed in other simulated scenarios.

4.3 TMT experiment

[TODO: add the simulation for TMT experiment, batch vs plex?]

5 Sample size calculation and power analysis

The proposed approach adjusts for the underlying protein abundance in the PTM significance analysis, which corrects the confounding factor with a cost of increased uncertainty. We compared the required sample size with or without the adjustment to highlight the property (Figure S10), based on a design of two conditions, in consideration of three pairs of standard deviations of log-intensities for modified and unmodified peptides: (0.2, 0.1), (0.2, 0.2), and (0.2, 0.4). We then characterized the advantage of general statistical modeling in complex designs in terms of sample size calculation (Figure S11) and power analysis (Figure S12).

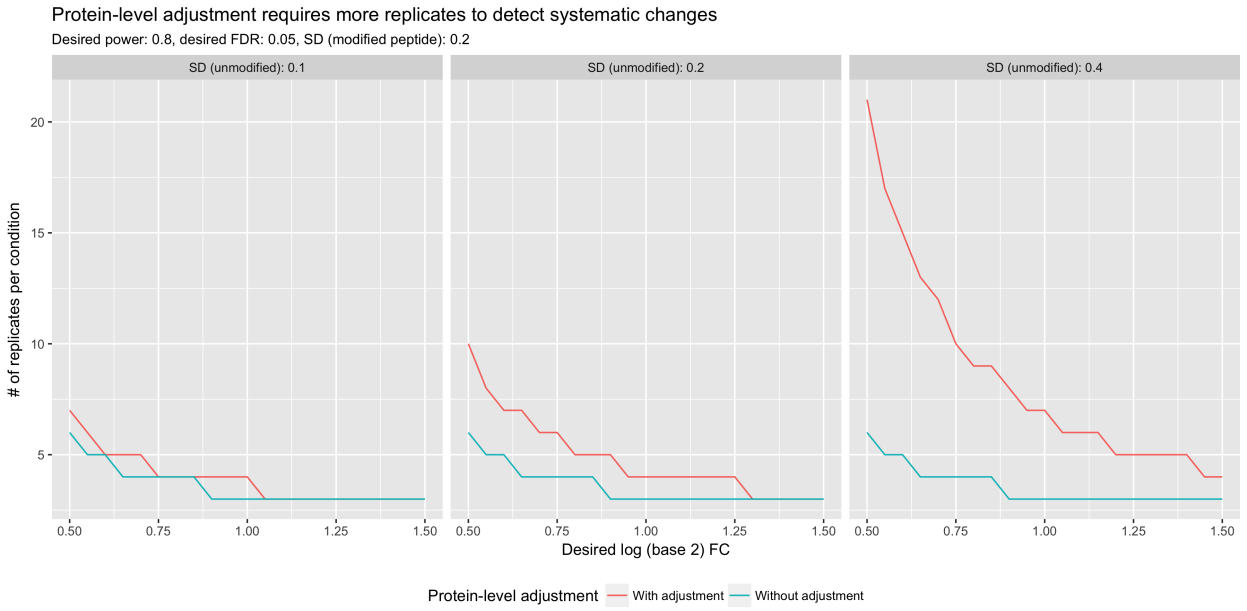

Figure S10: Protein-level adjustment relies on the inference of protein abundance, which introduces additional uncertainty in the estimate of PTM difference. Therefore, the required sample size to detect a systematic change is higher than as expected for standard differential analysis without adjustment. The discrepancy can be profound if the uncertainty associated with the protein abundance estimate is greater than that of PTM abundance estimate. Sample size calculation without accounting for the uncertainty would lead to under-powered studies.

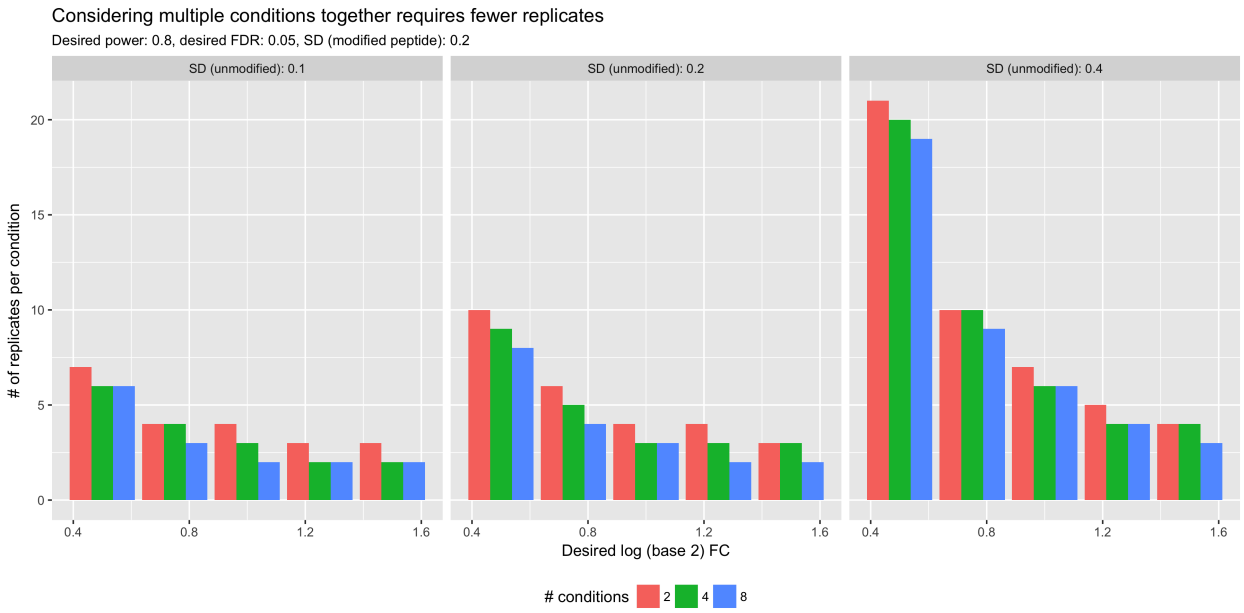

Figure S11: In complex designs, simultaneously analyzing all the conditions effectively increases the degrees of freedom and requires fewer replicates.

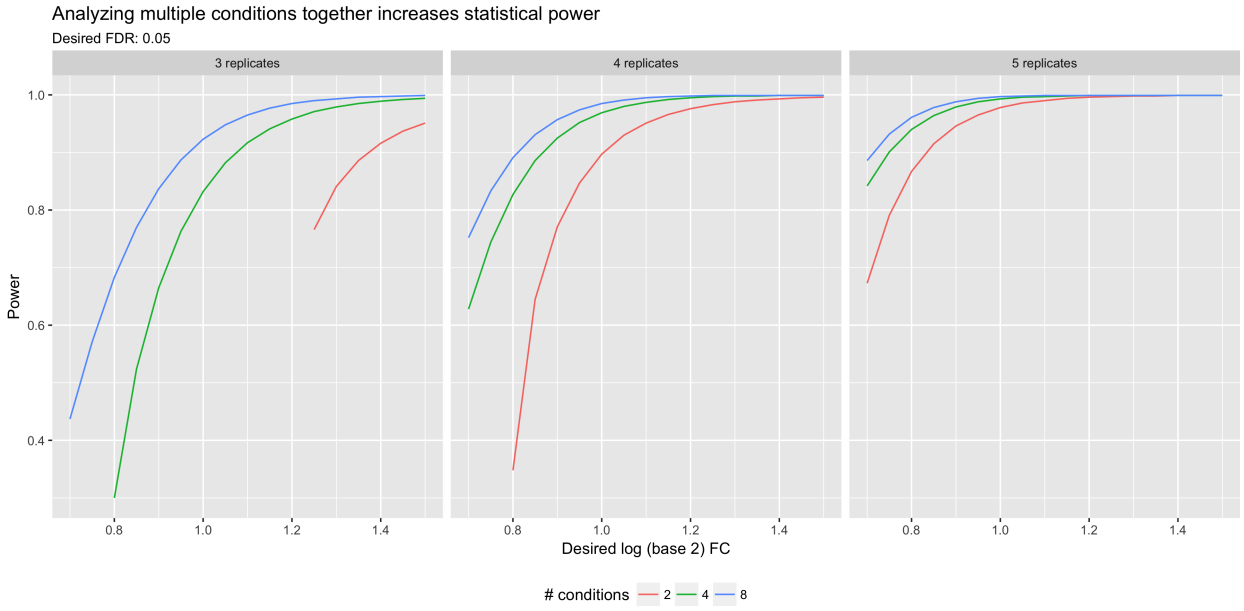

Figure S12: Statistical power can be improved by increasing the sample size and analyzing multiple conditions together.

6 Datasets : Biological investigation

[TODO: description and results]

References

1. Meena Choi, Ching-Yun Chang, Timothy Clough, Daniel Broudy, Trevor Killeen, Brendan MacLean, and Olga Vitek. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*, 30(17):2524–2526, 2014.

2. Michael H. Kutner, John Neter, Christopher J. Nachtsheim, and William Li. *Applied Linear Statistical Models*. McGraw-Hill/Irwin, 5 edition, 2004.

3. Ann L. Oberg and Olga Vitek. Statistical design of quantitative mass spectrometry-based proteomic experiments. *Journal of Proteome Research*, 8(5):2144–2156, 2009.