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

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

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

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

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

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

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

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

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

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

3.4.2 TMT experiment

The methods described in this paper can be applied to Tandem Mass Tag (TMT) experiments targeting PTMs, in addition to batch effects. The same hypothesis testing is performed to detect differences in adjusted log-abundance of modifications in Condition i and i', where the null hypothesis is that there is no difference in PTM abundance between 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 TMT experiments there is an additional source of variation coming from the mixtures used. This increases the complexity as there are many options for allocating samples to mixtures, including multiple mixtures, biological replicates, and technical replicates. Going forward we will denote the log-intensity of a modified peptide for Technical Replicate j of Mixture m for Condition i by y_{ijm}, where $m = 1, ..., M$. This is done in the same way for the unmodified peptide y^*_{ijm}.

The model assumes the variability of the mixture is a random effect and is not convoluted with the other variables in the model. The model assumes that the difference between conditions is the same between mixtures. For a given modified peptide and corresponding global protein a mixed effects model is fit as follows.

$$Y_{mtcb} = \mu + Mixture_m + Techrep(Mixture)_{jm} + Condition_c + Subject_{mcb} + \epsilon_{mtcb}$$
$$Y^*_{mtcb} = \mu^* + Mixture^*_m + Techrep(Mixture)^*_{jm} + Condition^*_c + Subject^*_{mcb} + \epsilon^*_{mtcb}$$

$$\text{Where } Mixture_m \sim N(0, \sigma^2_m), Techrep(Mixture)_{jm} \sim N(0, \sigma^2_{jm}), Subject \sim N(0, \sigma^2_s),$$
$$\epsilon_{mtcb} \sim N(0, sigma^2)$$

The log-fold change for adjusted PTM abundance is estimated as the average over mixtures and technical replicates

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

and the standard error of the estimate is

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

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

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

4 Computer simulation

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

4.1 Protein-level adjustment

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 or Limma taking as input the ratio between feature intensities of modified and unmodified peptides 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) proposed approach without adjusting for unmodified peptides c) t-test (with adjustment), d) t-test (no adjustment), e) Limma (with adjustment), and f) Limme (no adjustment).

In experiments of complex designs, multiple inter-related conditions are often compared together. The proposed approach and Limma leverage measurements in all conditions for the inference of the underlying abundance, whereas t-test uses measurements from the two conditions being compared. To highlight this distinction, multiple conditions of data were generated. The simulation was based on the following parameters:

4.1.1 Simulation 1

In the first simulation an experiment with many observations per protein/PTM and no missing data was created. This is not representative of a real experiment, but provides a baseline for model performance.

- 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, 0.75, 1.5, 2.25
- Difference in protein abundance between conditions: 0, 0.75, 1.5, 2.25
- Number of replicates: 2, 3, 5, 10
- Number of conditions: 2, 3, 4
- Number of realizations: 1000
- Number of features per PTM: 10
- Number of features per protein: 10

- Missing data: no missing value

The results are summarized from Figure **??** to Figure S5, including false positive rate with or without changes in protein abundance by the considered methods (Figure **??**), false positive rate, and overall accuracy. Additionally the results are compared to simulation 2 described in the next section.

When simulating the data, 500 of the 1000 peptides were simulated with a differencial fold change between conditons, while the other 500 were not. In terms of the 500 peptides that were not differencial, 250 were simulated with a change in both PTM and overall protein abundance, such that the change in PTM abundance is entirely due to the change in the overall protein. The other 250 were simulated with no significant change in either PTM or overall protein abundance. In this way there were 500 true positives and 500 true negative PTMs. These were used to calculate the summary statistics seen in the following plots.

4.1.2 Simulation 2

In the second simulation we introduced limited feature observations per PTM as well as masking a portion of the observation to simulate missing values. This is more in line with what we would expect with a real life experiment and provides a more realistic expectation of model performance.

- 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, 0.75, 1.5, 2.25
- Difference in protein abundance between conditions: 0, 0.75, 1.5, 2.25
- Number of replicates: 2, 3, 5, 10
- Number of conditions: 2, 3, 4
- Number of realizations: 1000
- Number of features per PTM: 2
- Number of features per protein: 10
- Missing data: 20% of the observations for PTMs and Proteins were masked with NA at random

The results are summarized from Figure S2 to Figure S5, including false positive rate with or without changes in protein abundance by the considered methods (Figure S2), false positive rate, and overall accuracy. Again the results are compared to simulation 1.

The portion of significant peptides were done in the same was as described in simulation 1.

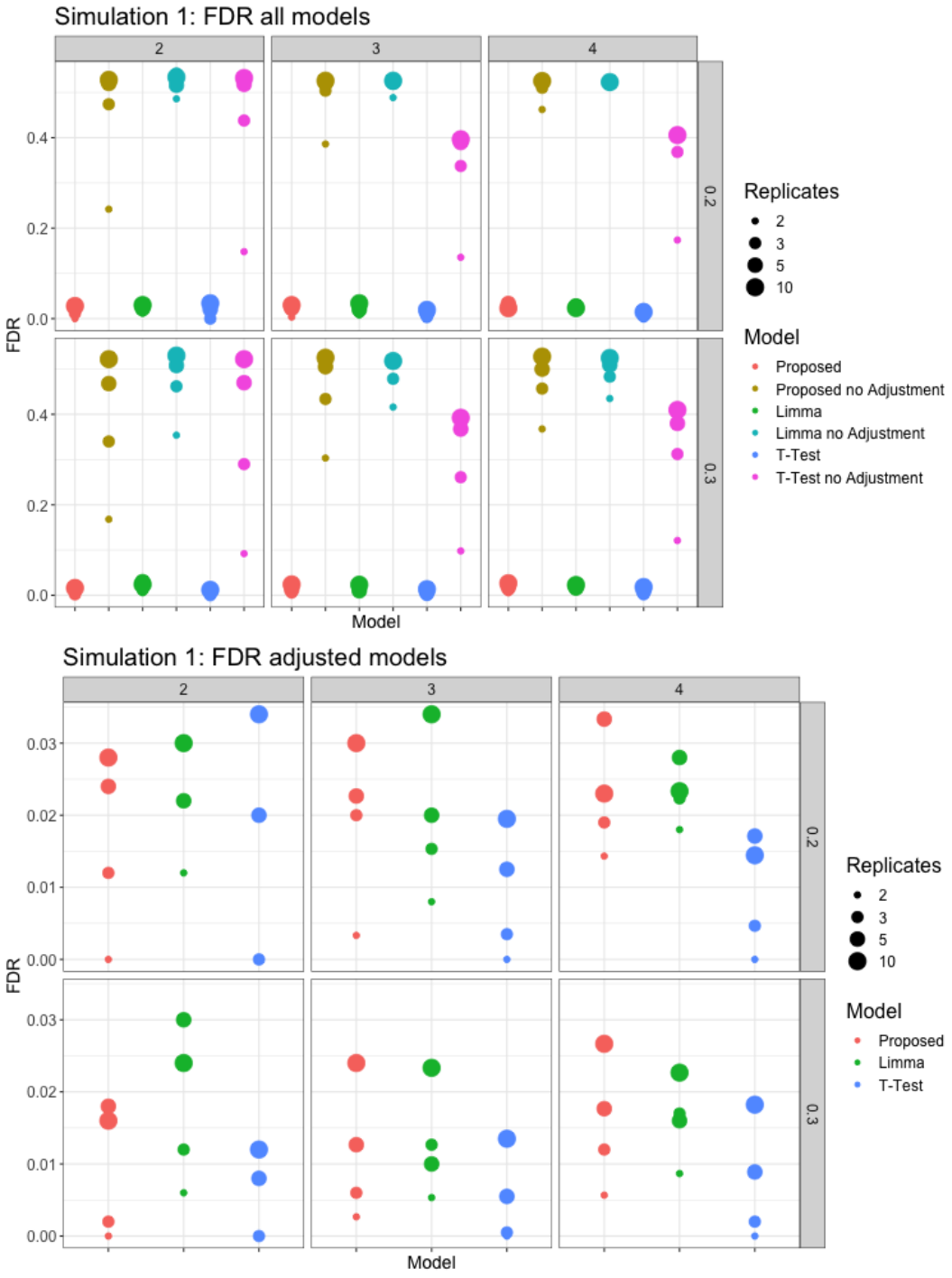

Figure S2: (upper plot)All the considered methods in simulation 1 correctly calibrated the Type I error rate when adjusting for changes in protein abundance. In comparison, the methods without accounting for the protein-level changes resulted in off-target, high false positive rates. (lower plot) The considered methods with protein adjustment are compared in detail. The proposed method generally performed better than Limma, but had a higher FDR than T-test. All methods performed strongly, with all FDR rates near or below .03.

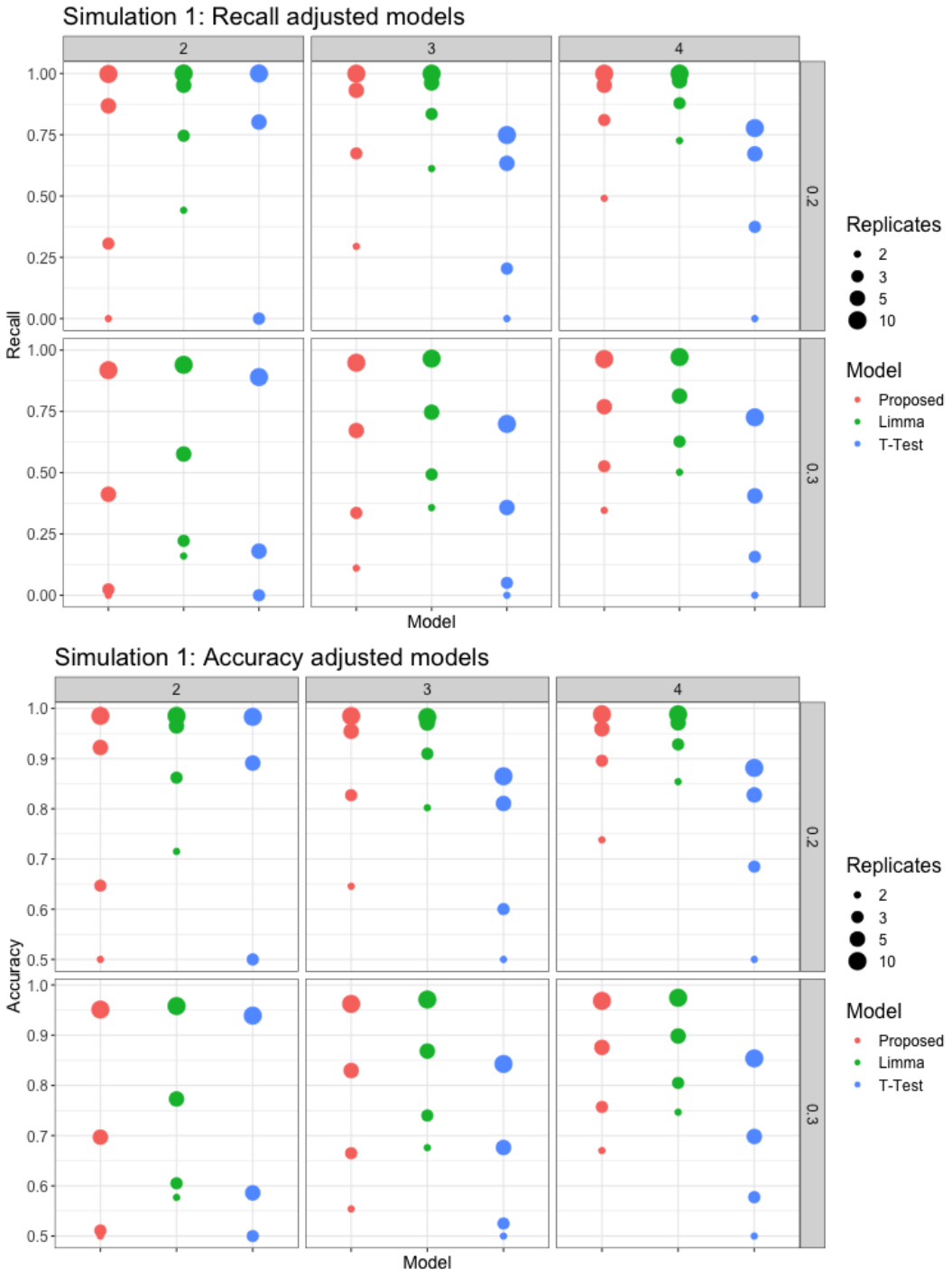

Figure S3: (upper plot) The recall for simulation 1 methods with adjustment were compared. It is clear that Limma performs the strongest here when the number of replicates were low. At higher replicates the performance of the proposed methods and Limma are comparable. T-test clearly performs worse across all methods. (lower plot) The overall accuracy plot mimics the observations in the recall plot. Limma performs stronger than the proposed method at lower replicates, while at higher they are comparable.

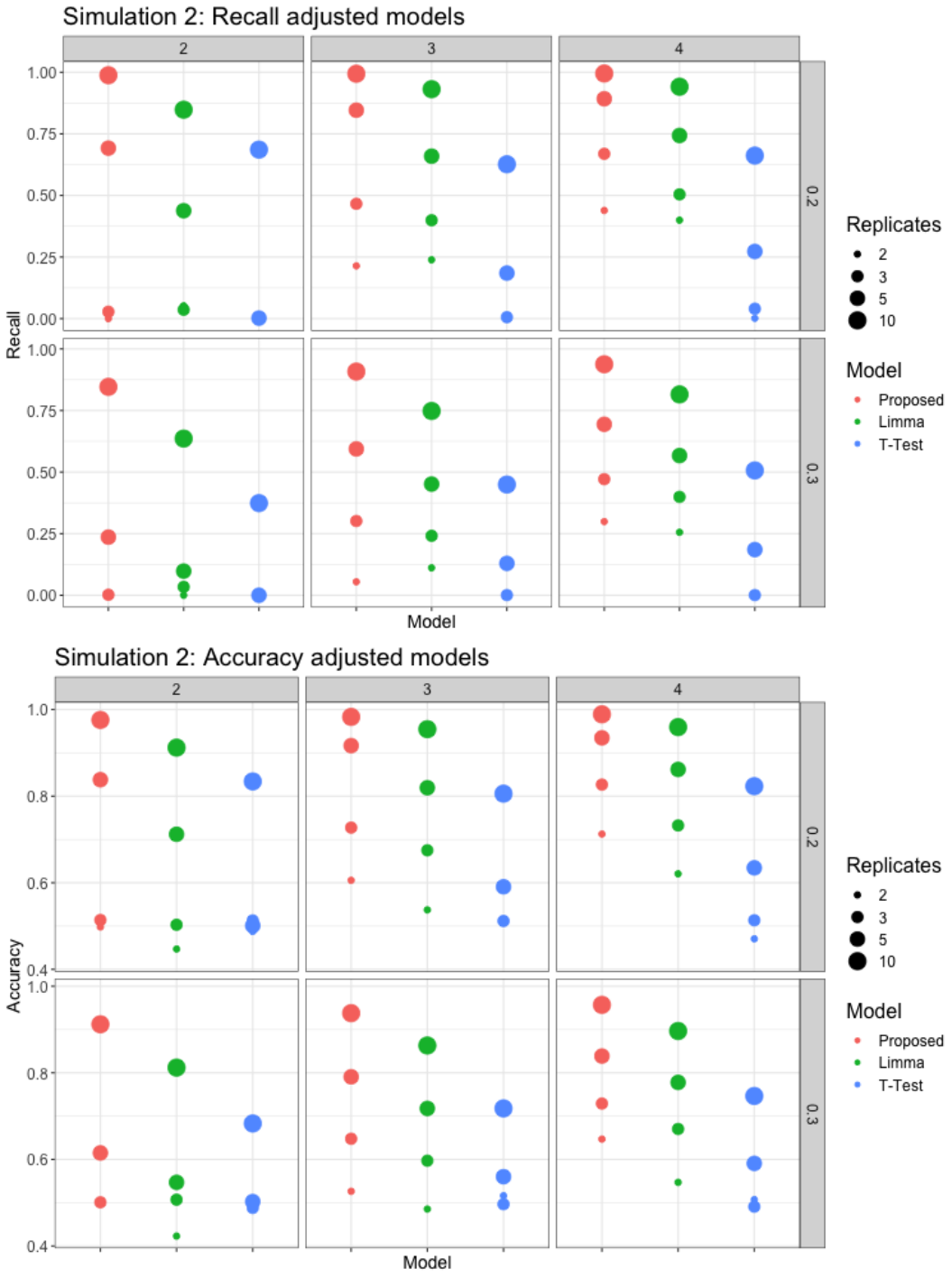

Figure S4: The advantage of using the proposed approach over Limma and *t*-test is apparent when looking at simulation 2, which includes limited observations and the presence of missing values. (upper plot) In the case of recall the proposed method performs stronger than limma and t-test in nearly every model. Even at lower replicates the proposed method still outperformed limma, especially when there were more conditions in the simulation. Again the lowest performance was observed using the *t*-test method. (lower plot) Again in the overall accuracy plots, the results are similar to recall. The propsed method performed the strongest across all methods, even when replicates were low. While Limma shows strong performance in a clean experiment, when real world data problems are introduced it is clear the proposed method is more robust.

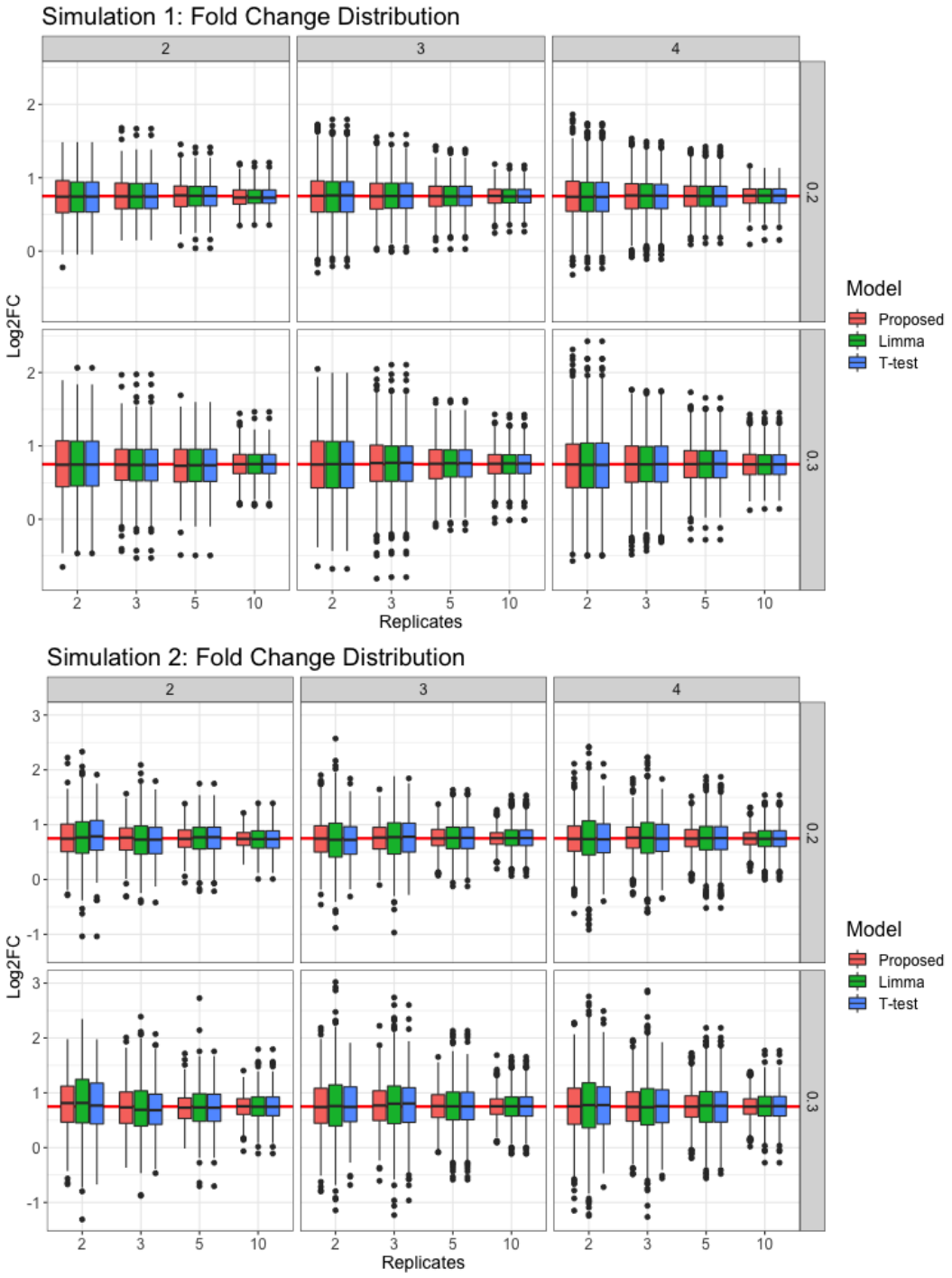

Figure S5: (upper plot) In simulation 1 all considered methods correctly estimated the fold change between conditions, with a median fold change estimation of .75 in all methods. The distributions around the median were also consistent acorss all methods. Predictably the distribution quantiles decrease as the number of replicates increases. (lower plot) In simulation 2 again all methods correctly estimated the fold change with a median log change of .75. The proposed method in this simulation had a visibly tighter distribution around the median. Both Limma and t-test showed a wider range around the fold change. In this case the proposed method showed a stronger performed in correctly estimating the fold change for all peptides.

4.2 Benchmark experiment

A custom designed experiment with labeling was used to assess the performance of the proposed method in a real experimental setting. Heavy-labeld KGG modified peptides were used as spike-in peptides. The spike-in peptides were mixed with human lysate to create four mixture conditions. Two sets of data were acquired for each mixture: KGG enriched + LC-MS, and LC-MS only. The KGG enriched dataset included the spike-in peptides, as well as modified and unmodified human lysate. The LC-MS dataset included only unmodified peptides. The spike-in peptides where a significant fold change between conditions are expected to be differencial, whereas none of the human lysate peptides in any comparison are expected to be differencial.

Again we consider three different methods and assess their performance: the method proposed in this paper, Limma, and two sample t-test. All methods are analyzed after adjusting for changes in overall protein level. The propsed method summarizes feature intensities up to the run level using Tukey's Median Polish, while the other methods use the log of summed feature intensities. The results are summarized from Figure S6 to Figure S10, including volcano plots, model summary statsitics, and fold change analysis.

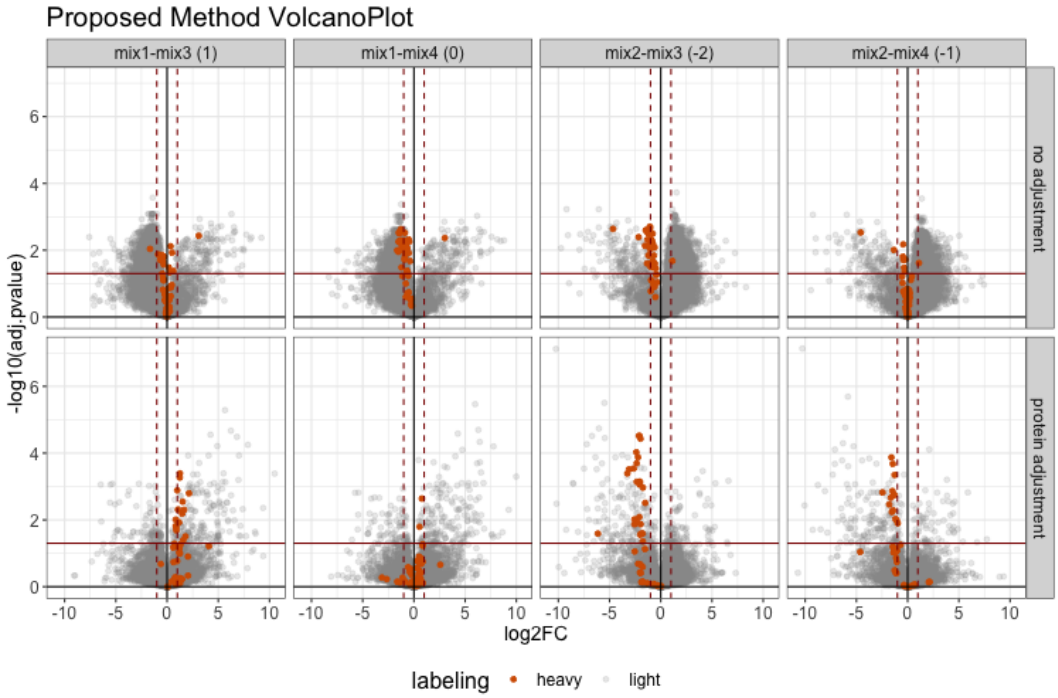

Figure S6: Using the proposed method to model the benchmark experiment the spike in peptides (colored red) do not follow the expected log fold change before adjustment. After adjusting for changes in overall protein abundance the spike in peptides are more in line with expectation. Additionally the background grey colored peptides showed many false positives before adjustment. After adjustment the false positives were decreased considerably.

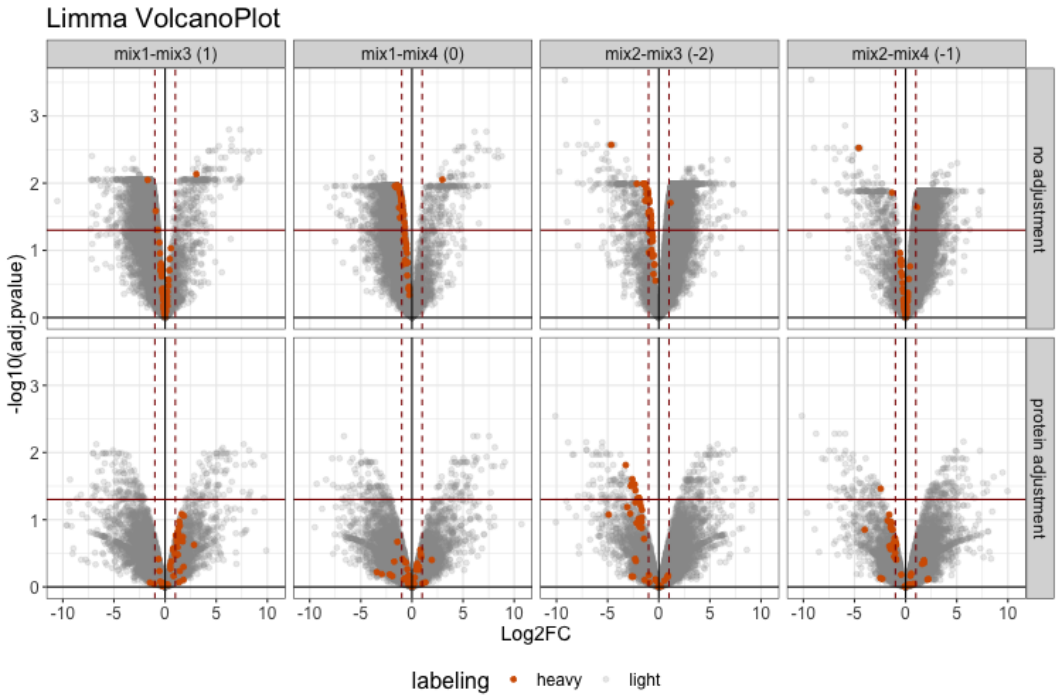

Figure S7: Using the limma method the spike in peptides follow the expected log fold change better after adjusting for changes in protein level. However, while the fold change is much more accurate, the majority of spike in peptides do not have a significante adjusted pvalue. In terms of false positives, the results are very similar to the propsed method, with many false positives before adjustment and much fewer afterward.

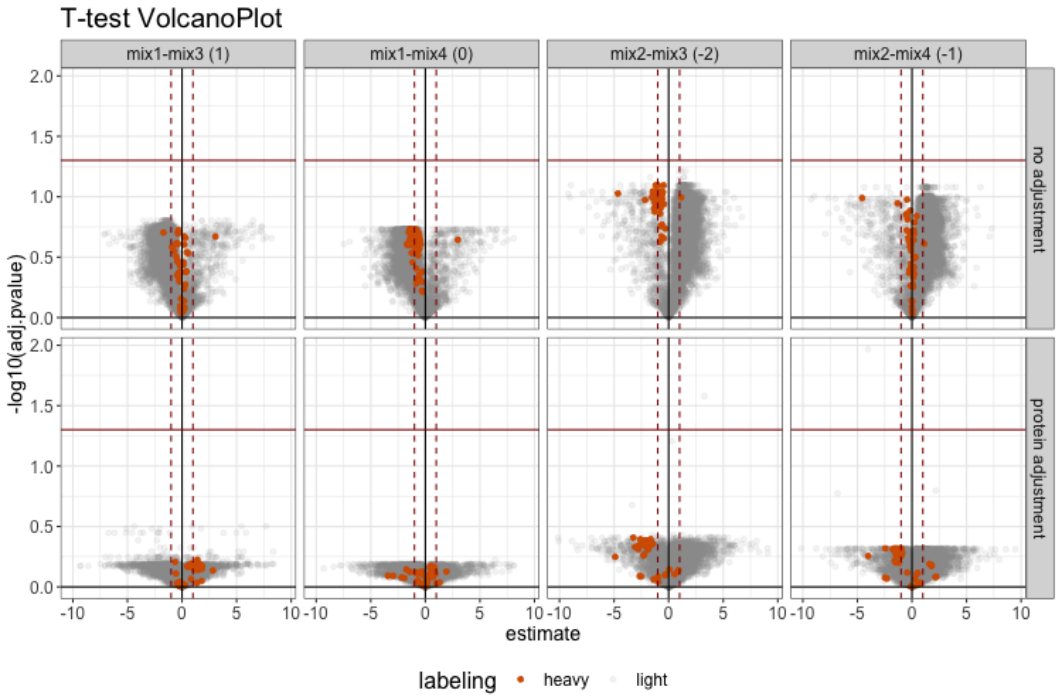

Figure S8: Using the two sample t-test, none of the comparisons either before or after adjustment do not show any significant peptides. With that being said, the fold change of the spike in peptides is much closer to expectation after adjusting for global protein abundance.

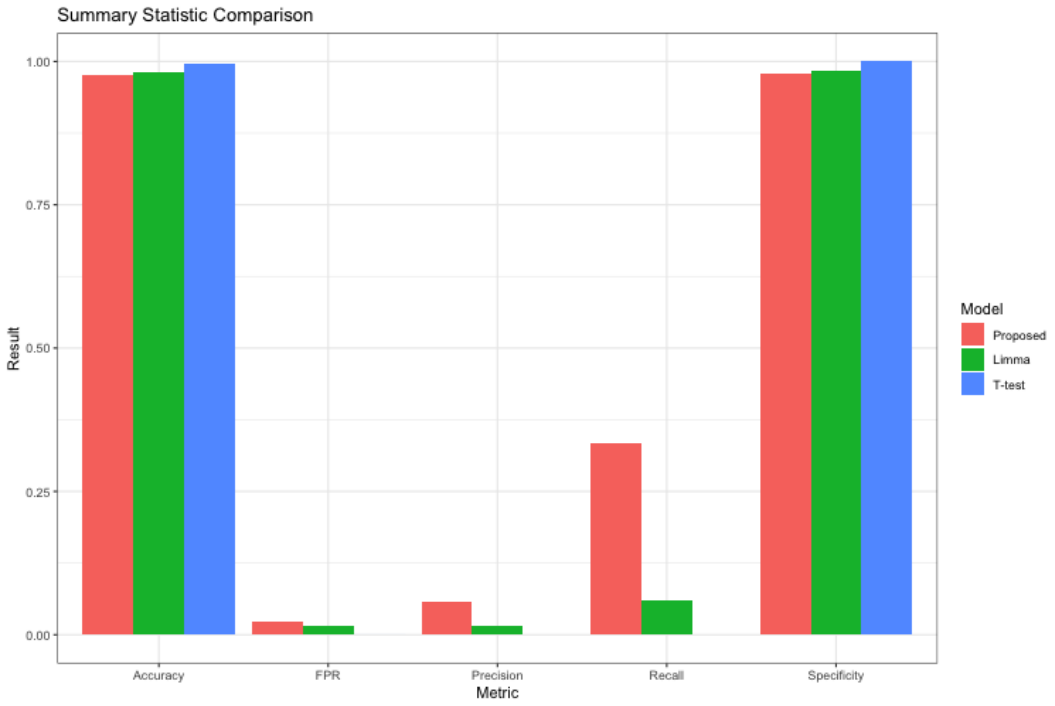

Figure S9: Comparing the summary statistics between methods it is clear that the proposed method performs the strongest. In terms of accuracy and specificity the three methods are close, with Limma and t-test showing slightly higher values. Accuracy and specificity are dominated by the large number of true negatives (background peptides) compared to the true positives (spike in peptides). In terms of recall, the proposed approach far out performed the other two methods, showing that it correctly labeled the most spike in peptides.

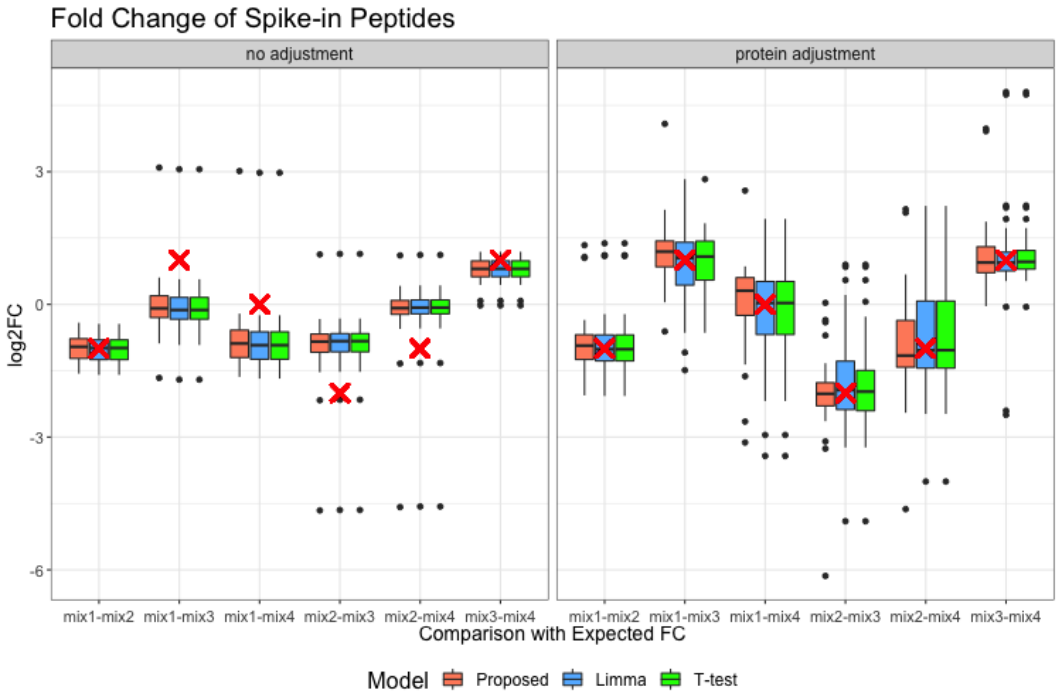

Figure S10: The fold change misalignment before adjustment is illustrated in the boxplots. The true log fold change is indicated by the red 'X' marks. Before adjustment the boxplots are not close to the true fold change. After adjustment all three models generally have median near the true fold change. In terms of model comparison, the proposed method shows a much tighter distribution around the true fold change, whole both Limma and t-test are much wider.

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 variation. We compared the required sample size with or without the adjustment to highlight the property (Figure S11), 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 S12) and power analysis (Figure S13).

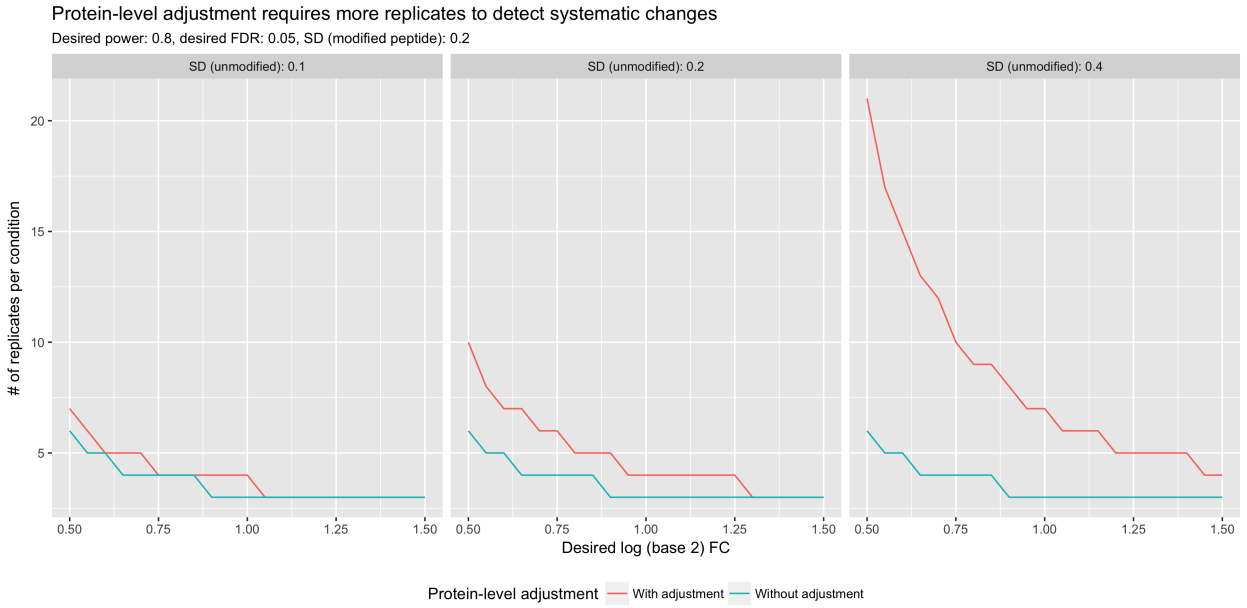

Figure S11: Protein-level adjustment relies on the inference of protein abundance, which introduces additional uncertainty in the estimate of PTM difference. Therefore, the required sample size to detect a systematic change is higher than as expected for standard differential analysis without adjustment. 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.

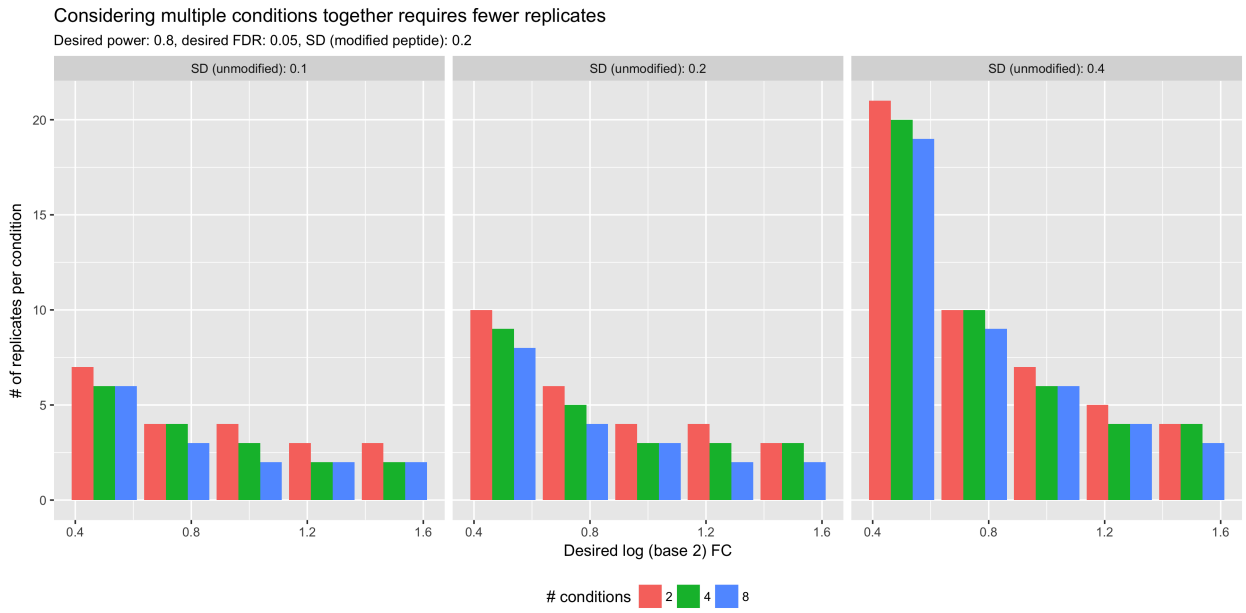

Figure S12: In complex designs, simultaneously analyzing all the conditions effectively increases the degrees of freedom and requires fewer replicates.

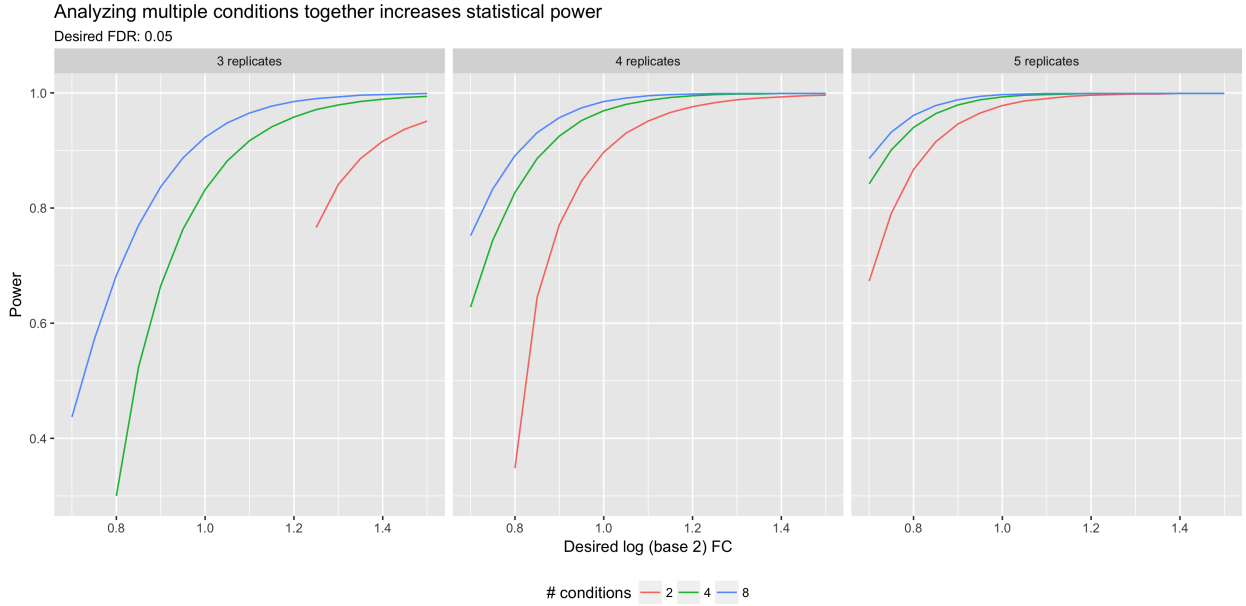

Figure S13: Statistical power can be improved by increasing the sample size and analyzing multiple conditions together.

6 Datasets : Biological investigation

6.1 USP30

This experiment was analyzed using the proposed method, limma, and two sample t-test. Becuase this is a real biological experiment we do not know the true postives and negatives as with the spike-in and simulation evaluations. The results of each method can be seen in Figure **??**

Model	Before Adjustment	After Adjustment
Proposed	4502	1625
Limma	1211	361
t-test	28	7

Figure S14.a: Significant modified peptides per model

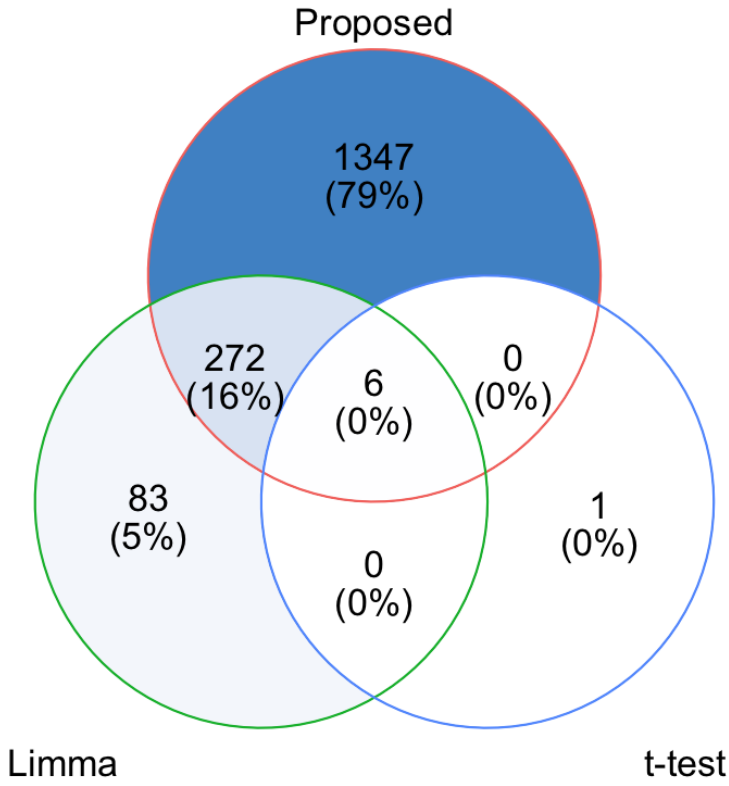

Figure S14.b: Shared significant modified peptides between models

Figure S14: A. The proposed method identifies far more signficant modified peptides compared to Limma and *t*-test. All methods identify substantially fewer peptides after adjusting for changes in global protein level. B. The proposed method and Limma share 16% of all peptides identified as significant. The proposed method is the only model that identified 79% of the peptides, while Limma identified 5% that no other model did. *t*-test only identified 7 peptides as signififcant, with all but 1 being shared with the other two models.

7 Further Considerations

7.0.1 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 **??**.

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.