MSstatsPTM: an R/Bioconductor software for detecting quantitative changes in post-translational modifications

Devon Kohler[1]*, Tsung-Heng Tsai[2]*, Ting Huang[1], Erik Verschueren[4], Trent Hinkle[3], Meena Choi[3]**, Olga Vitek[1]**

[1]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
[2]Kent State University, Kent, OH, USA
[3]Genentech, South San Francisco, CA, USA
[4]Galapagos, Mechelen, Antwerp, Belgium

Overview: The scientific community widely utilizes mass spectrometry (MS)-based proteomics to quantify the abundance of proteins and their post-translational modifications (PTMs). Experiments targeting PTMs face several specific challenges. These include the low abundance, few representative peptides, and convolution with abundance changes in the overall protein expression. Due to these challenges, a robust approach to estimate relative changes in PTMs should combine PTM sites over several peptides, replicates in multiple conditions, and consider sources of confounding present in the experiment. We propose a general statistical model and workflow that is both reproducible and comprehensive. The method measures PTM and protein abundance by summarizing intensities through Tukey's median polish method. Then a model based on the family of linear mixed-effects models is fit. Finally, the PTM abundances are adjusted to remove variance from changes in the overall protein. The package can handle a diverse range of acquisition types, including label-free, DDA, DIA, and TMT. We implement this model in the free and open-source R package **MSstatsPTM** and available at the links below.

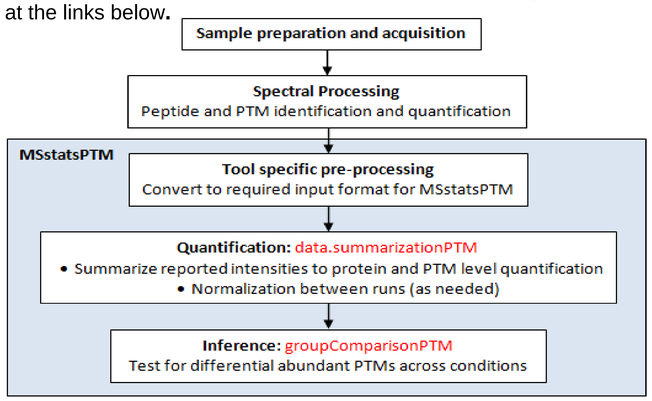

Figure 1: Workflow of MSstatsPTM.

1. Sample preparation and acquisition

Two experiments were analyzed in order to demonstrate MSstatsPTM's ability to model experiments generated via many different labeling methods.

The first experiment was custom designed to provide a benchmark. Heavy-labeled 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.

	No change in spike-in	Decrease in spike-in
No change in human lysate	**Mix 1** 2xA 2xB	**Mix 2** 2xA 1xB
Decrease in human lysate	**Mix 3** 1xA 2xB	**Mix 4** 1xA 1xB

Figure 2: Benchmark experimental design

The TMT experiment targeted primary murine macrophages infected with Shigella flexneri (S.flexneri). Tandem mass tagging was used to analyze differences in proteins, phosphorylation, and ubiquitination. The TMT acquisition was run with two mixtures consisting of 11 samples. Wild type (WT) controls were compared to ATG16L1-deficient BMDMs (cKO) over an extended time period.

Package Link(s):

1. https://github.com/tsunghengtsai/MSstatsPTM/tree/master/R
2. https://github.com/Vitek-Lab/MSstatsTMTPTM

Manual(s):

1. https://bioconductor.org/packages/release/bioc/manuals/MSstatsPTM/man/MSstatsPTM.pdf
2. https://bioconductor.org/packages/release/bioc/manuals/MSstatsTMTPTM/man/MSstatsTMTPTM.pdf

2. Summarization methods

Run-level summarization of peptide ions (features) for each protein and PTM site is carried out as in the sub-plot model of MSstats and MSstatsTMT (depending on type of experiment). This involves both imputation of missing values and summarization of feature intensities using Tukey's median polish.

Label-Free Summarization

		Condition 1				...	Condition C			
		Biorep 1	Biorep 1	...	Biorep B		Biorep 1	Biorep 1	...	Biorep B
		Run1	Run2	...	Run J	...	Run1	Run2	...	Run J
Modified	Feature 1	y	y	...	NA	...	y	NA	...	y
	Feature 2	y	NA	...	y	...	y	y	...	y
	...	...	...	...	...	...	...	...	...	...
	Feature K	y	y	...	NA	...	y	NA	...	y
Unmodified	Feature 1	y	y	...	y	...	NA	NA	...	NA
	Feature 2	NA	y	...	y	...	y	NA	...	y
	...	...	...	...	...	...	...	...	...	...
	Feature L	y	NA	...	NA	...	NA	y	...	y

After summarization

	Condition 1				...	Condition C			
	Sub	Sub	...	Sub	...	Sub	Sub	...	Sub
Modified	μ	μ	...	μ	...	μ	μ	...	μ
Unmodified	μ*	μ*	...	μ*	...	μ*	μ*	...	μ*

Figure 3: The input data from a label-free experiment for MSstatsPTM for one modification and protein. There are two separate measurements for both the modified and unmodified peptide with their corresponding features (peptide ions). Each cell denotes the log2 reporter ion intensity of the observed feature and NA indicates a missing value.

3. Modeling Methods

Both the protein and PTM sites are fit with linear models. The PTM model is then adjusted to remove confounding of the global protein level.

PTM Model (Label-Free)

$$\mu_{sc} = \psi + Condition_c + Subject_{cs} + \epsilon_{cs}$$

Protein Model (Label-Free)

$$\mu^*_{sc} = \psi^* + Condition^*_c + Subject^*_{cs} + \epsilon^*_{cs}$$

$$\text{where } \textstyle\sum_{c=1}^{C} Condition_c = 0, Subject_{cs} \overset{iid}{\sim} N(0, \sigma^2_S),\ \epsilon_{cs} \overset{iid}{\sim} N(0, \sigma^2)$$

4. Adjustment Method

Hypothesis	$H_0 : \Delta = (\mu_c - \mu_{c'}) - \left(\mu^*_c - \mu^*_{c'}\right) = 0$ $H_a : \Delta = (\mu_c - \mu_{c'}) - \left(\mu^*_c - \mu^*_{c'}\right) \neq 0$
Log-fold change	$\hat{\Delta} = \left[\frac{1}{J}\left(\hat{y}_{c+} - \hat{y}_{c'+}\right)\right] - \left[\frac{1}{J}\left(\hat{y}^*_{c+} - \hat{y}^*_{c'+}\right)\right]$
Standard Error	$\left[\frac{2}{J}\left(\hat{\sigma}^2_\gamma + \hat{\sigma}^2_{\gamma^*}\right)\right]^{1/2}$
Degrees of Freedom	$\left(\hat{\sigma}^2_\gamma + \hat{\sigma}^2_{\gamma^*}\right)^2 \Big/ \left(\frac{\hat{\sigma}^4_\gamma}{\mathrm{df}(\gamma)} + \frac{\hat{\sigma}^4_{\gamma^*}}{\mathrm{df}(\gamma^*)}\right)$

Figure 4: The formulas used to adjust PTM for overall protein level. The log-fold change of adjusted PTM, Δ, includes the summarization over features. The standard error and degrees of freedom of Δ are also shown above.

5. Result: Benchmark Experiment

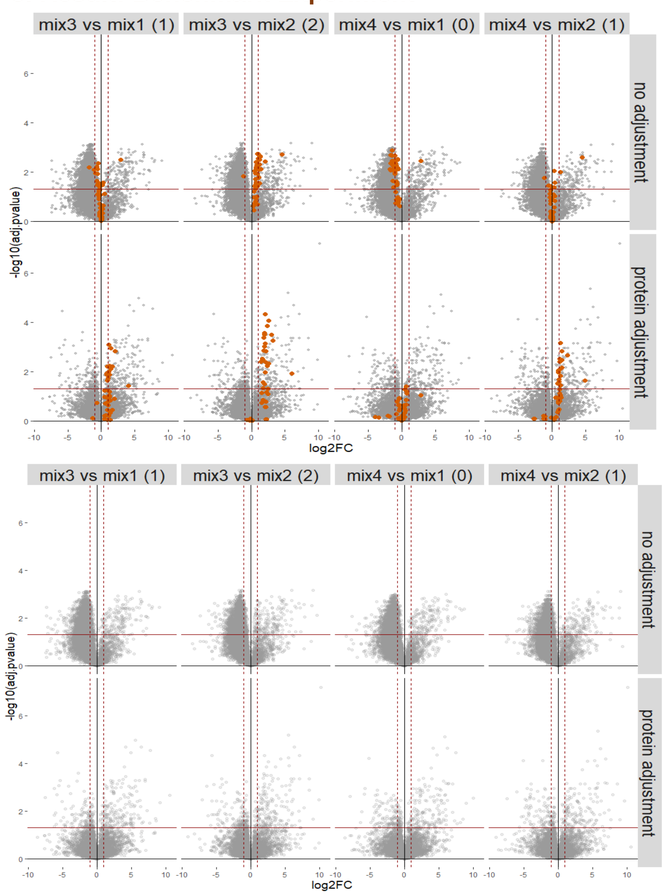

Figure 5: Significance analysis of spike-in peptides before and after adjusting for changes in protein level. The spike-in peptides are colored red in the plot. The expected Fold Change is listed next to the title of each subplot. The spike-in peptides follow expectation only after adjustment is applied.

Figure 6: Light labeled modified peptides before after adjusting for overall protein level. Before adjustment there were many false positives. When protein level adjustment is applied, the number of false positives decreases dramatically.

6. Result: TMT Experiment

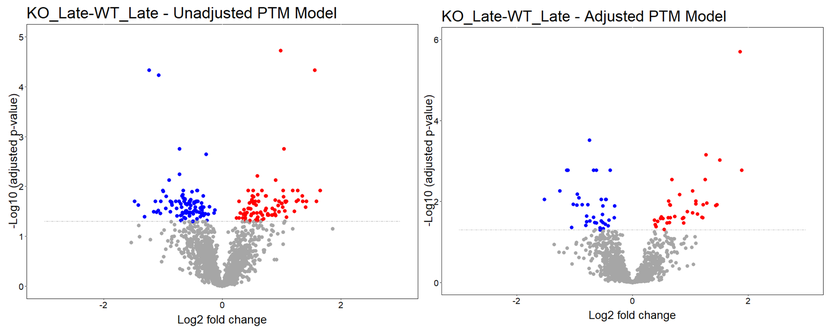

Figure 8: Significance analysis of PTMs in the comparison WT-KO before and after adjusting for total protein abundance. The number of significant modifications decreased after adjustment.

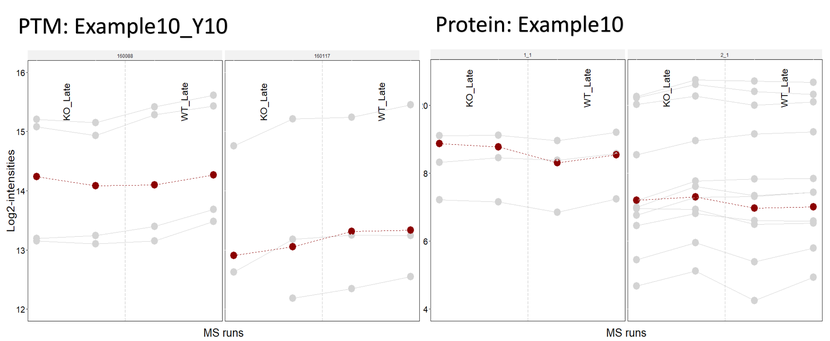

Figure 7: Profile plot of an example protein with a modification at site Y10 for the comparison between the cKO and WT observed 3 hours after injection. The relative log2-intensity of the overall protein is higher in cKO than WT. In contrast, the modification shows a higher log2-intensity in WT and lower in cKO.