SpikeIn MSstatsPTM Method Comparison

Devon Kohler

3/7/2022

```
library(MSstatsPTM)
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.1.2
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## Warning: package 'tibble' was built under R version 4.1.2

## Warning: package 'tidyr' was built under R version 4.1.2

## Warning: package 'readr' was built under R version 4.1.2

## Warning: package 'dplyr' was built under R version 4.1.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x purrr::transpose() masks data.table::transpose()
```

```
library(limma)
```

```
## Warning: package 'limma' was built under R version 4.1.1
```

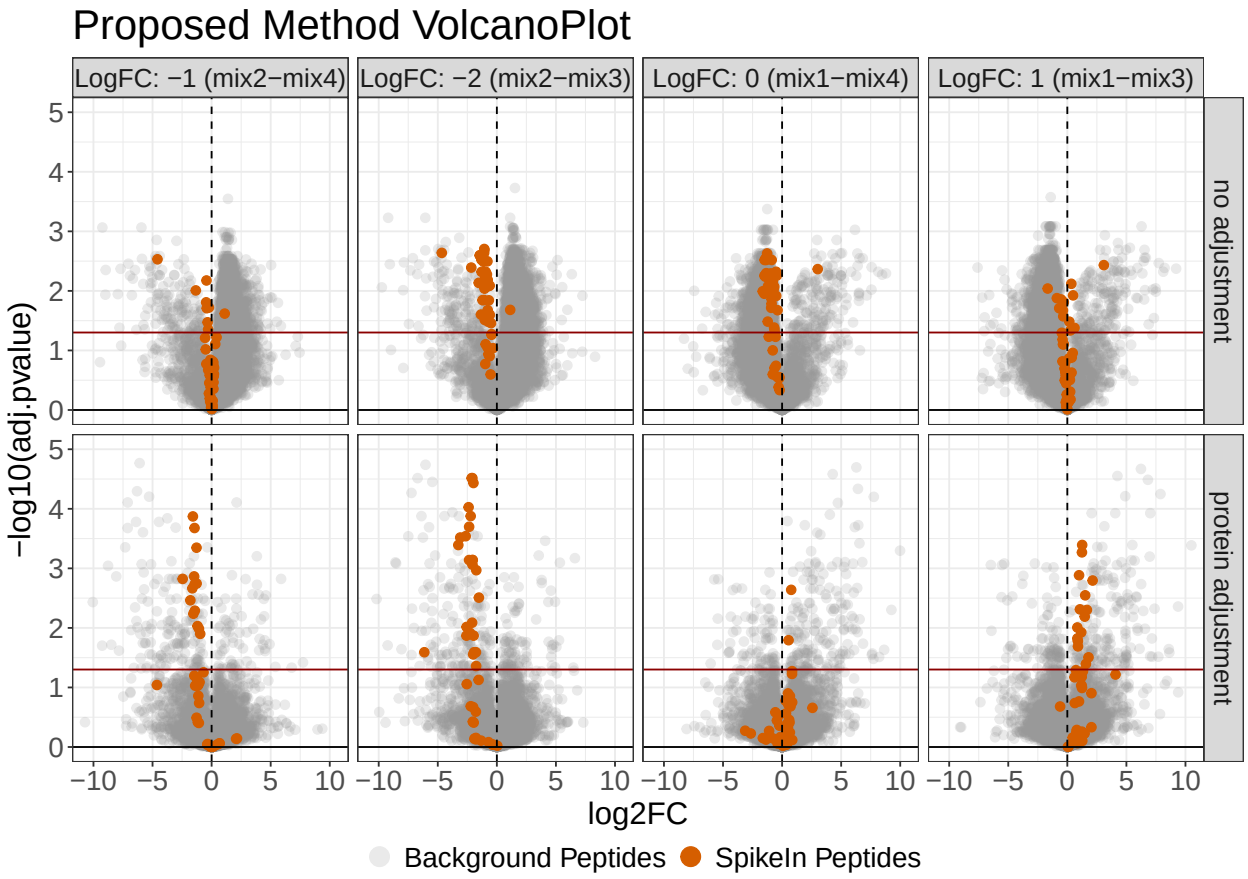
Spike-in Dataset Method Comparison

This markdown file examines the custom made spike-in analysis targeting PTMs. This allows us to know which PTMs are actually differential and which are not. The models used in this analysis are: MSstatsPTM, limma, and two condition t-test. For MSstatsPTM the full workflow is used for run-level summarization and modeling. Limma and t-test use feature averaging for run summarization and then their coresponding models are fit. The model results are shown first and then a comparison is made.

MSstatsPTM

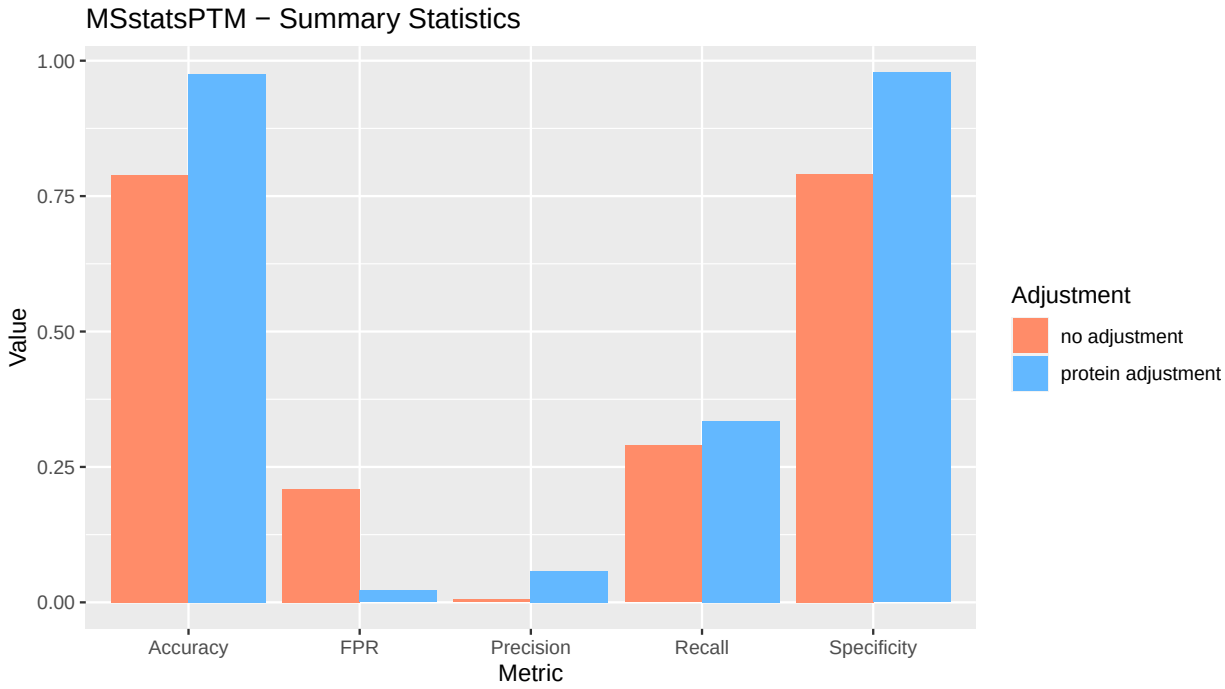
Volcano Plot First a comparison between the adjusted and unadjusted models are shown. The spike-in peptides are colored red and their expected log Fold Change is indicated in the subplot title. The background peptides are colored grey and should all be insignificant.

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



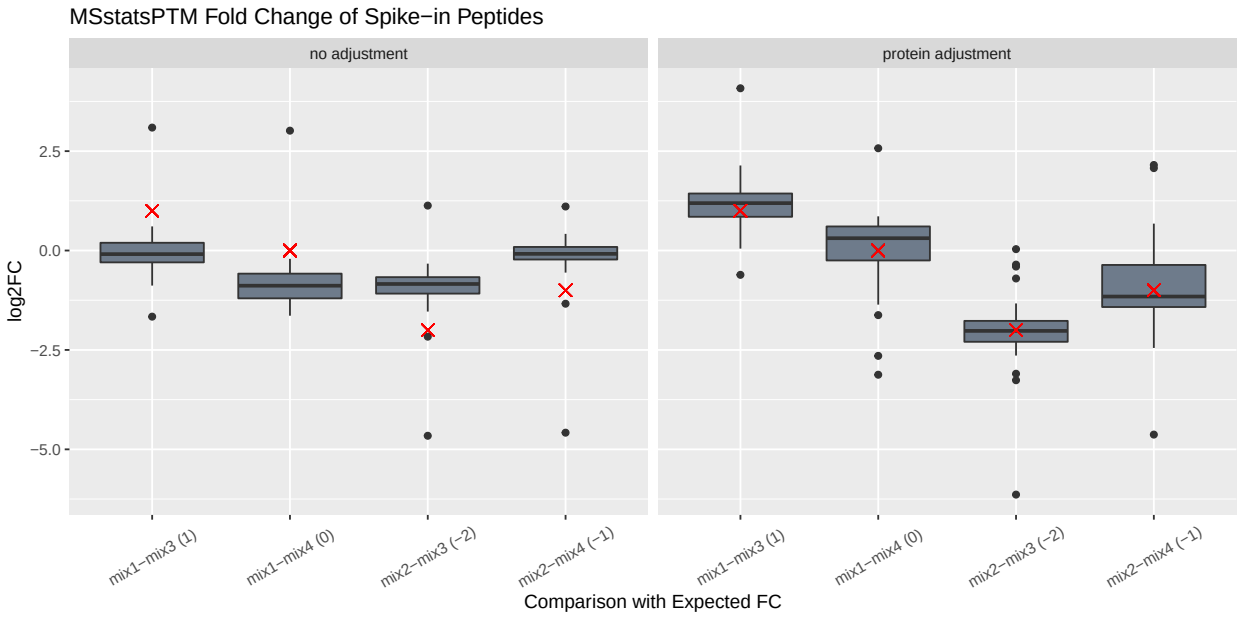
We can clearly see the effect of the adjustment in the volcano plots. Before adjustment, the red spike-in peptides did not follow the expected fold change. Additionally there are a large amount of false positives, shown as grey dots above the significance threshold. After applying protein level adjustment the red spike-in peptides are much more in line with expectation and there are many less false positives.

Summary Statistics Summary statistics for the models are given below.



In the summary statistics we can clearly see the advantage of the protein adjustment. Overall accuracy increases to near 100% (although we need to keep in mind it is dominated by the large number of true negatives). The False Positive Rate decreases substantially. Precision increases but is still quiet low. This is mainly due to the very large number of true negatives. While we were able to correctly categorize a large portion of the true negatives, there are still a large number that were missed and dominate this statistic. Recall only saw a slight increase in the adjusted model. This is because, while the fold change does now follow what we expected, a large number of the spike-in peptides still had an adjusted p value over .05. Finally specificity also saw a large increase when adding in the protein adjustment, indicating we are correctly identifying most of the background peptides as negatives.

Log Fold Change - Spike-in Peptides Lastly, the plot below looks at the distribution spike-in peptides log fold changes. The expected log Fold Change is indicated by the red `X` marks.



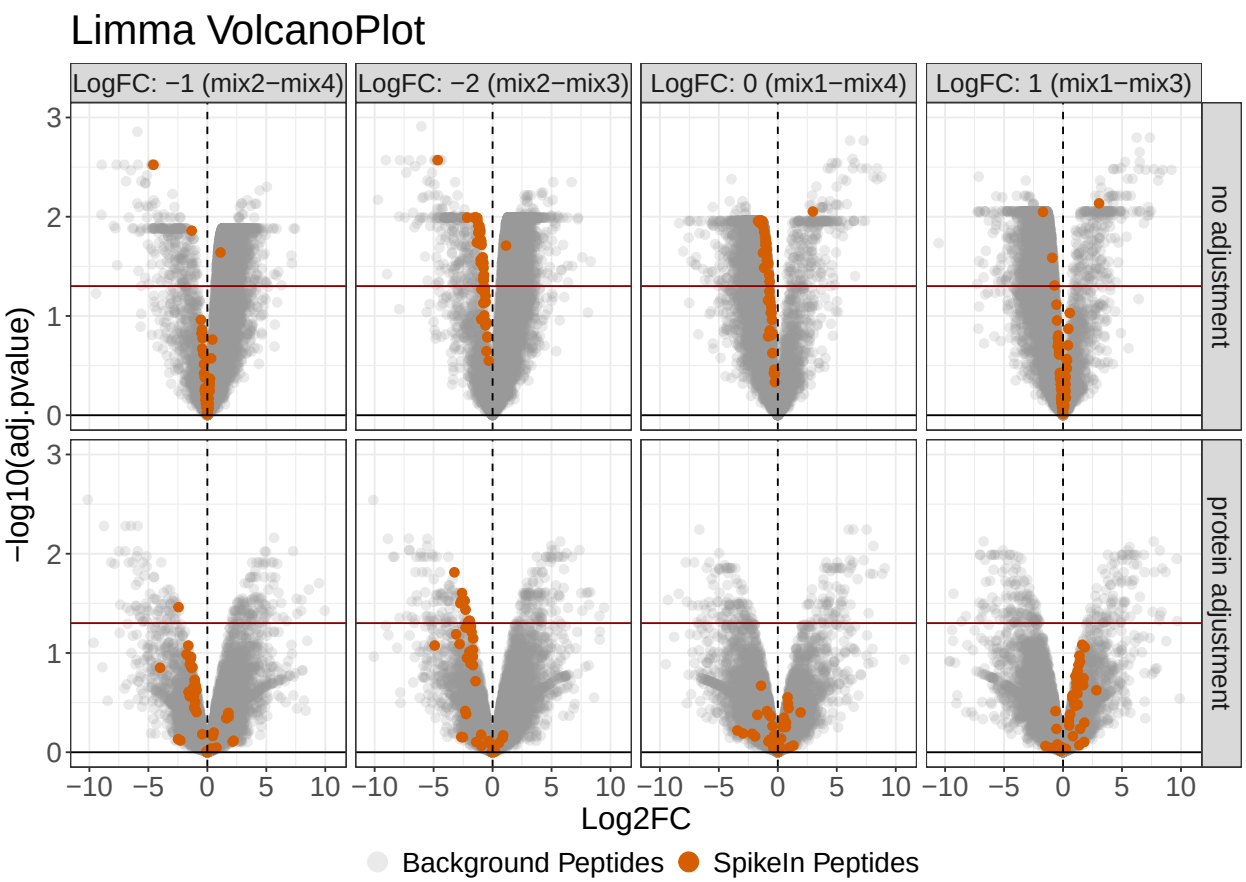
We can see in the plot above that the fold change of the spike-in peptides falls in line with expectation much better after adjustment. The expected change falls very close to the median value, and is at least within the 1st and 3rd quartile boxes in every comparison. In the unadjusted plot, the expected falls outside the quartile boxes in the majority of comparisons.

Limma

We will repeat the above analysis for Limma.

Volcano Plot First a comparison between the adjusted and unadjusted models are shown. Again the spike-in peptides are colored red and their expected log Fold Change is indicated in the subplot title. The background peptides are colored grey and should all be insignificant.

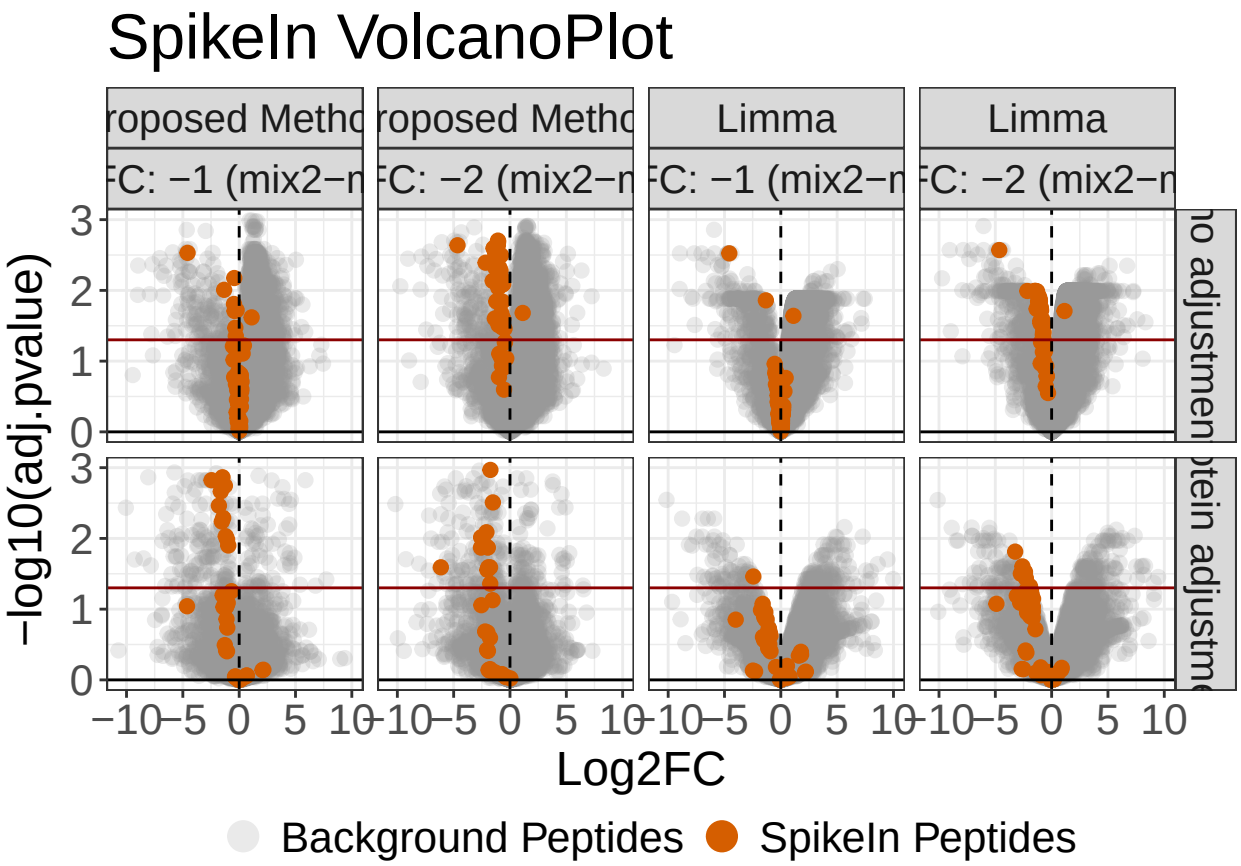
```
## Warning: Removed 2 rows containing missing values (geom_point).
```



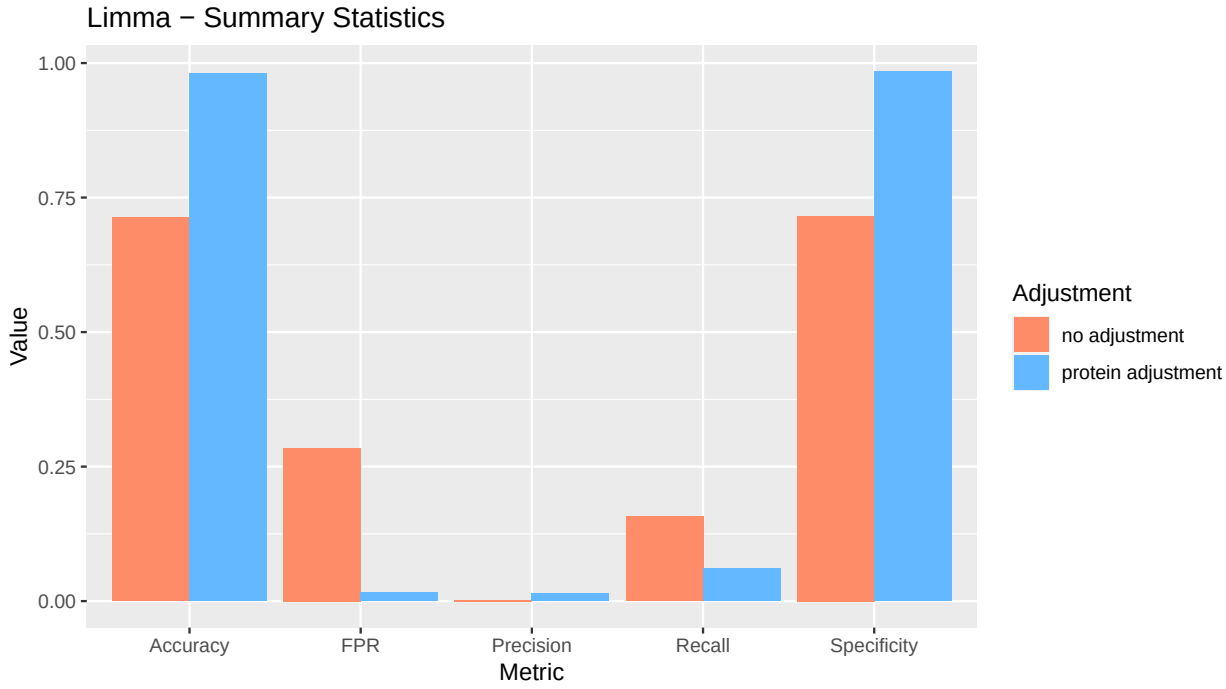
The limma model shows an improvement in terms of the false positives, but loses insight into the majority of spike-in peptides. After adjustment, the spike-in peptides are much more inline with the expected log2FC, however they are mostly below the significant pvalue threshold.

Create combo plot of proposed and limma

```
## Warning: Removed 85 rows containing missing values (geom_point).
```



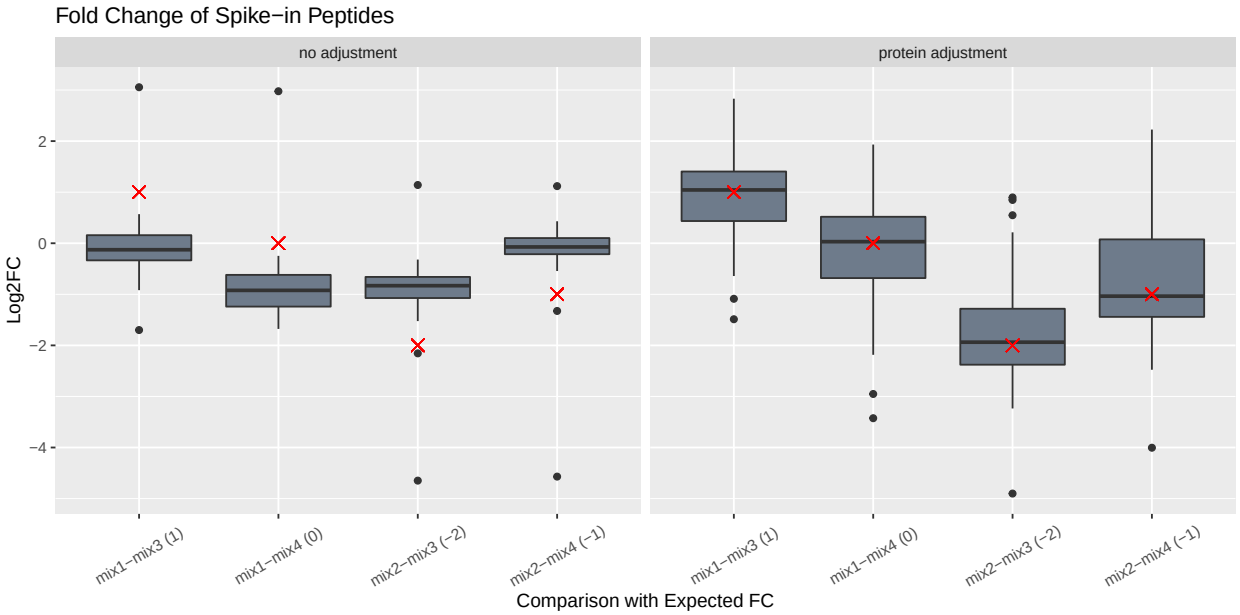
Summary Statistics Summary statistics for the models are given below.



The summary statistics repeat what we saw in the volcano plots. The accuracy and specificity are both

much higher after adjustment due to the decrease in false positives. Of course this is also seen in the lower FPR. Recall after adjustment is much lower. We are correctly labeling very few of the true positives after adjustment. This is in contrast to MSstatsPTM which actually saw an increase in recall after adjustment.

Log Fold Change - Spike-in Peptides Again the last thing we will look at is the distribution spike-in peptides log fold changes. The expected log Fold Change is indicated by the red `X` marks.



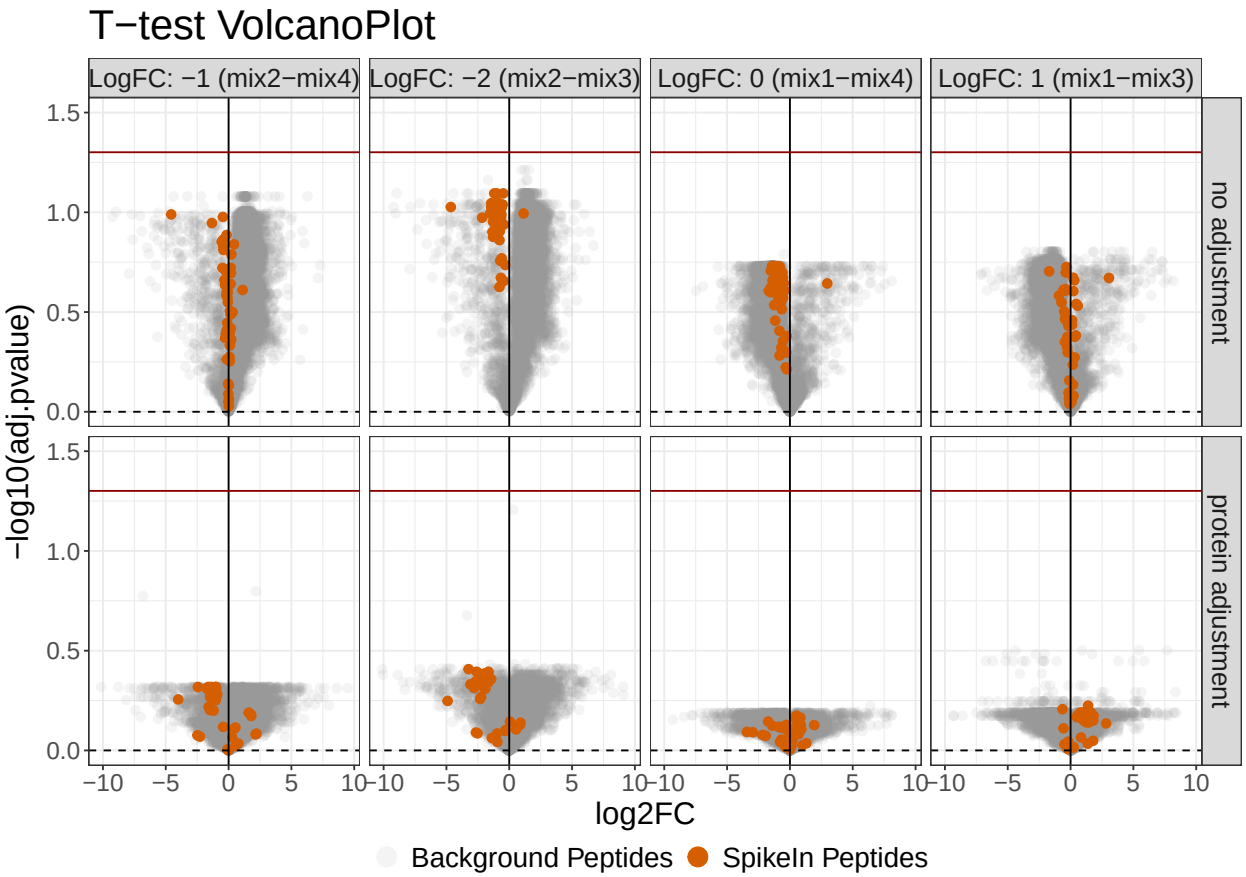
Once again we see a strong improvement in the Log2FC of spike-in peptides after adjustment. In contrast to MSstatsPTM the distribution after adjustment is much wider, with a median close to the true FC. This larger variance around the median indicates that while we are generally seeing the true FC, the results are not as strong as MSstatsPTM.

T-Test

Finally we will look at the t-test results.

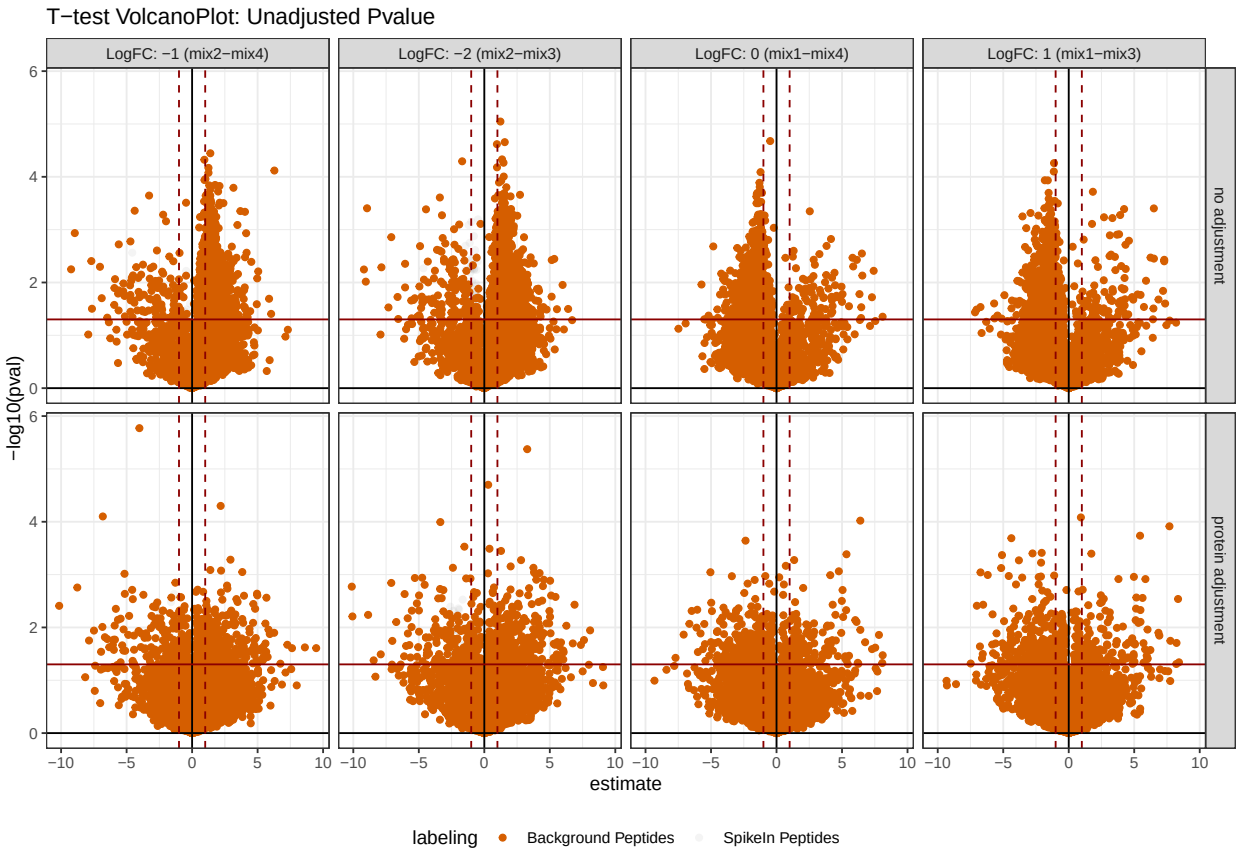
Volcano Plot Again we look at the volcano plots. The spike-in peptides are colored red and their expected log Fold Change is indicated in the subplot title. The background peptides are colored grey and should all be insignificant.

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



The results of the t-test are the weakest of the models we have observed. Both before and after adjustment we see no significant results. This may be due to an increased variance in the models, as each t-test model must be fit between two conditions and cannot share information. Additionally we only have two observations per condition, thus if one observation is missing the t-test model breaks down and the model cannot be fit.

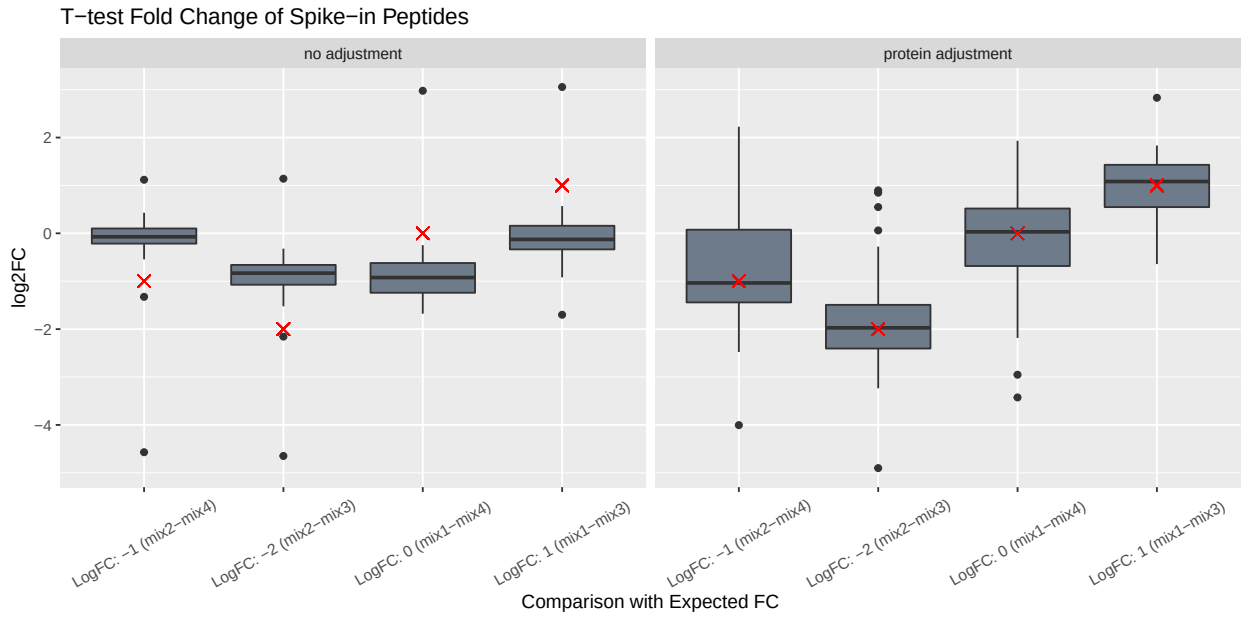
To get a closer look at the effect of the adjustment we will look at the volcano plots using the unadjusted pvalue.



The unadjusted pvalue volcano plots look very similar to what we have seen in the previous two models. The false positives decrease and the spike-in peptides follow the expected FC closer after adjustment.

Summary Statistics The summary statistics for this model are skipped because the adjusted pvalue shows no positive results.

Log Fold Change - Spike-in Peptides Lastly, we will again look at the distribution of the spike-in peptides log fold changes. The expected log Fold Change is indicated by the red `X` marks.

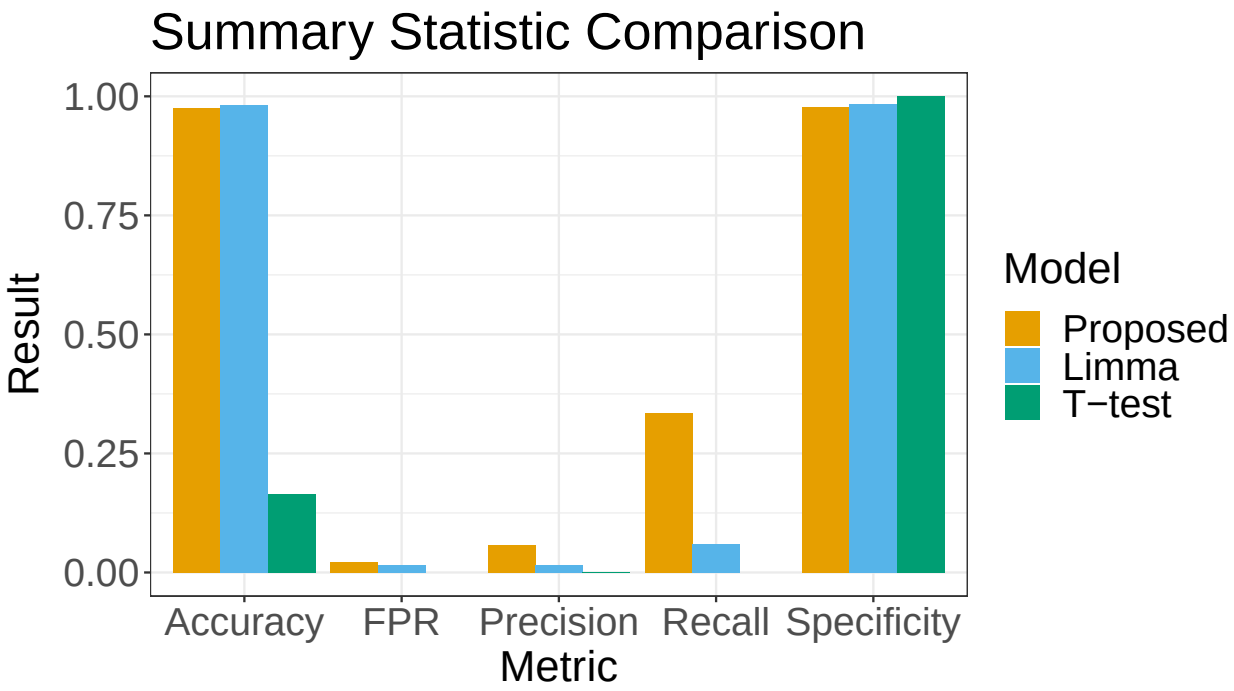


This plots is very similar to the limma plot. The spike-in peptides follow the expected FC much better after adjustment. However the effect is much wider around the true FC than MSstatsPTM, meaning the effect is less accurate.

Summary

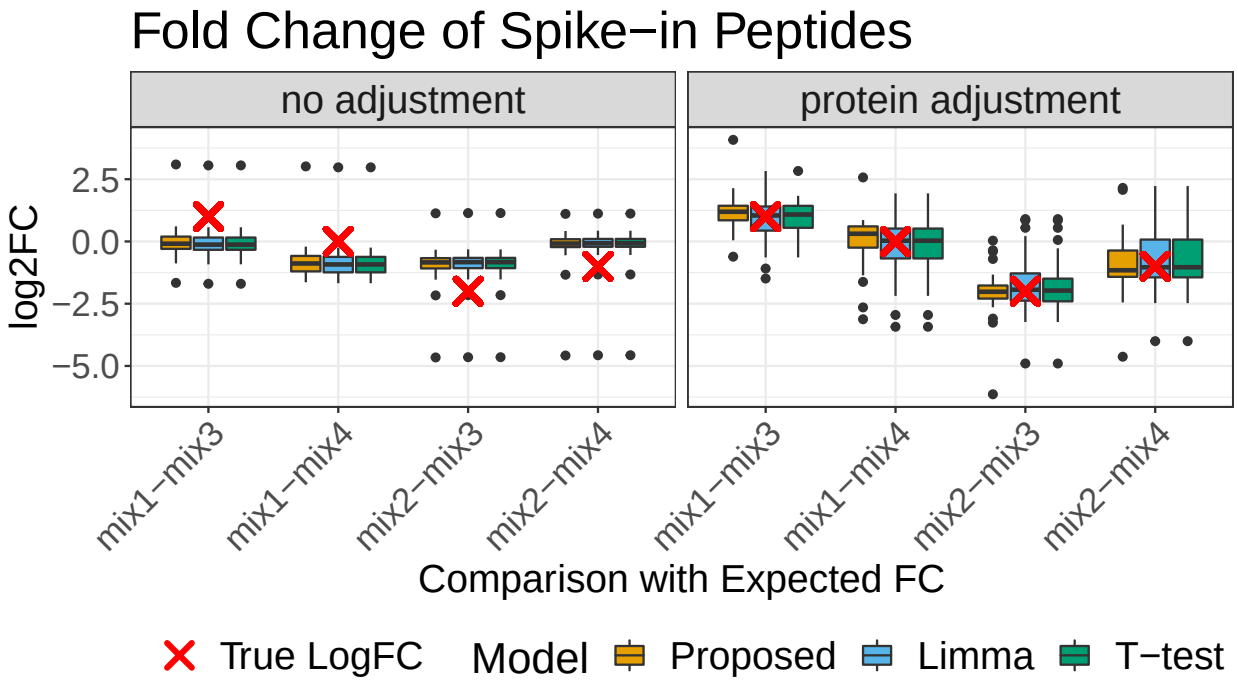
In the following plot we compare the summary statsitics of the adjusted models. Note T-test is shown but all values are zero.

```
## Warning: Removed 2 rows containing missing values (geom_col).
```



This plot confirms that MSstatsPTM and limma are comparable in terms of FDR, however MSstatsPTM performs much better when identifying the true positives, spike-in, peptides.

In the following plot we compare the log2FC of the spike-in peptides across models.



In this plot we can see that the unadjusted models are very similar and do not follow the expected FC. In comparison the adjusted models all generally follow the expectation, however MSstatsPTM has a tighter distribution around the expected FC. In contrast the other two models have a wider range around the true FC.