MSstatsPTM Simulation Analysis

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()
```

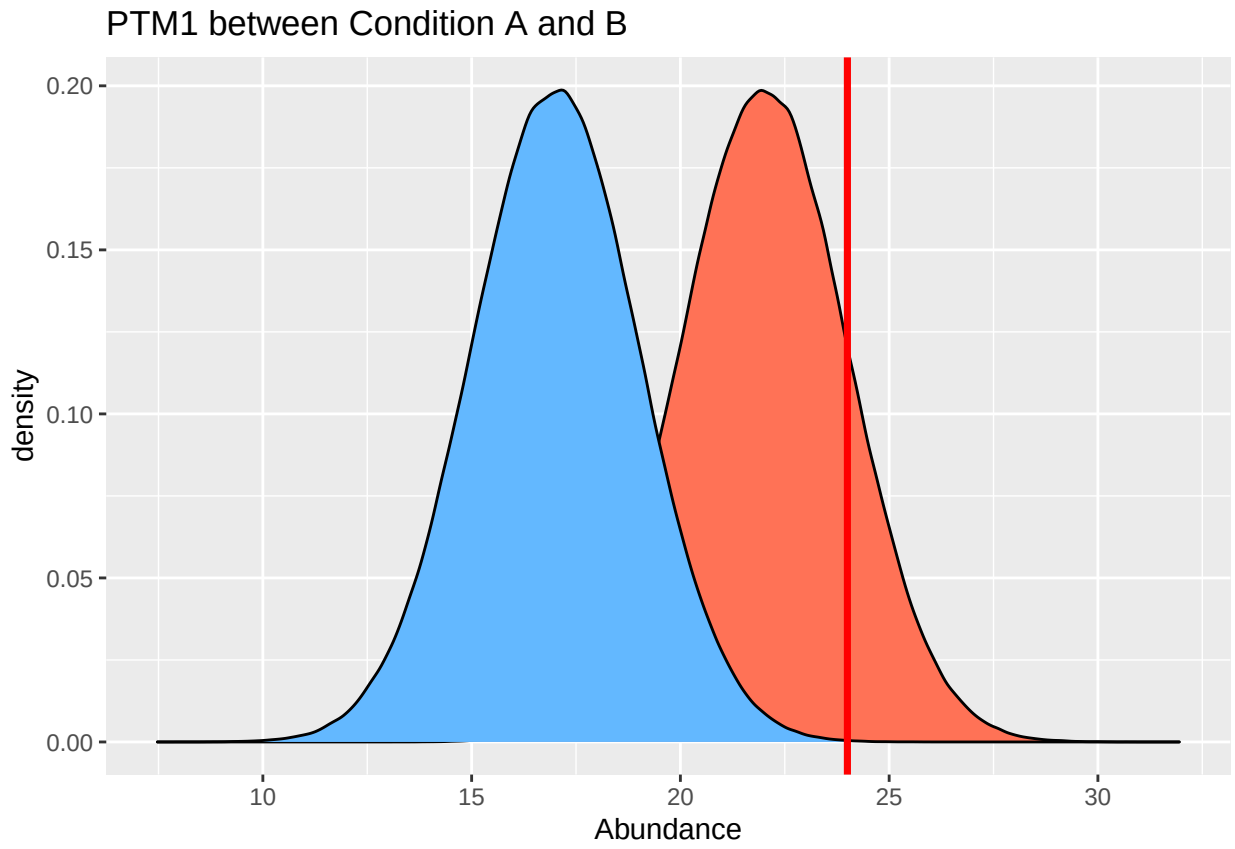
Computer Simulation

Simulation Methods

To simulate data the `PTMsimulateExperiment()` function was used. This function allows us to vary the number of conditions, replicates, number of proteins, number of sites per protein, number of spectral features per site/protein, mean log2-abundance of PTM and PROTEIN, deviation from the mean log2-abundance in each group, standard deviation among replicates, and standard deviation among log2-intensities.

Three different statistical modeling methods were applied to the simulated data: MSstatsPTM, limma, and two-sample t-test. These methods were tested both with and without applying protein level adjustment. MSstatsPTM uses TMP for summarization and post modeling calculations for the adjustment. To adjust limma and t-test, the run-level data was averaged for both ptm and protein datasets and then combined. The resulting dataset was then used for limma and t-test.

Graphs can help us visualize the components of the simulation.



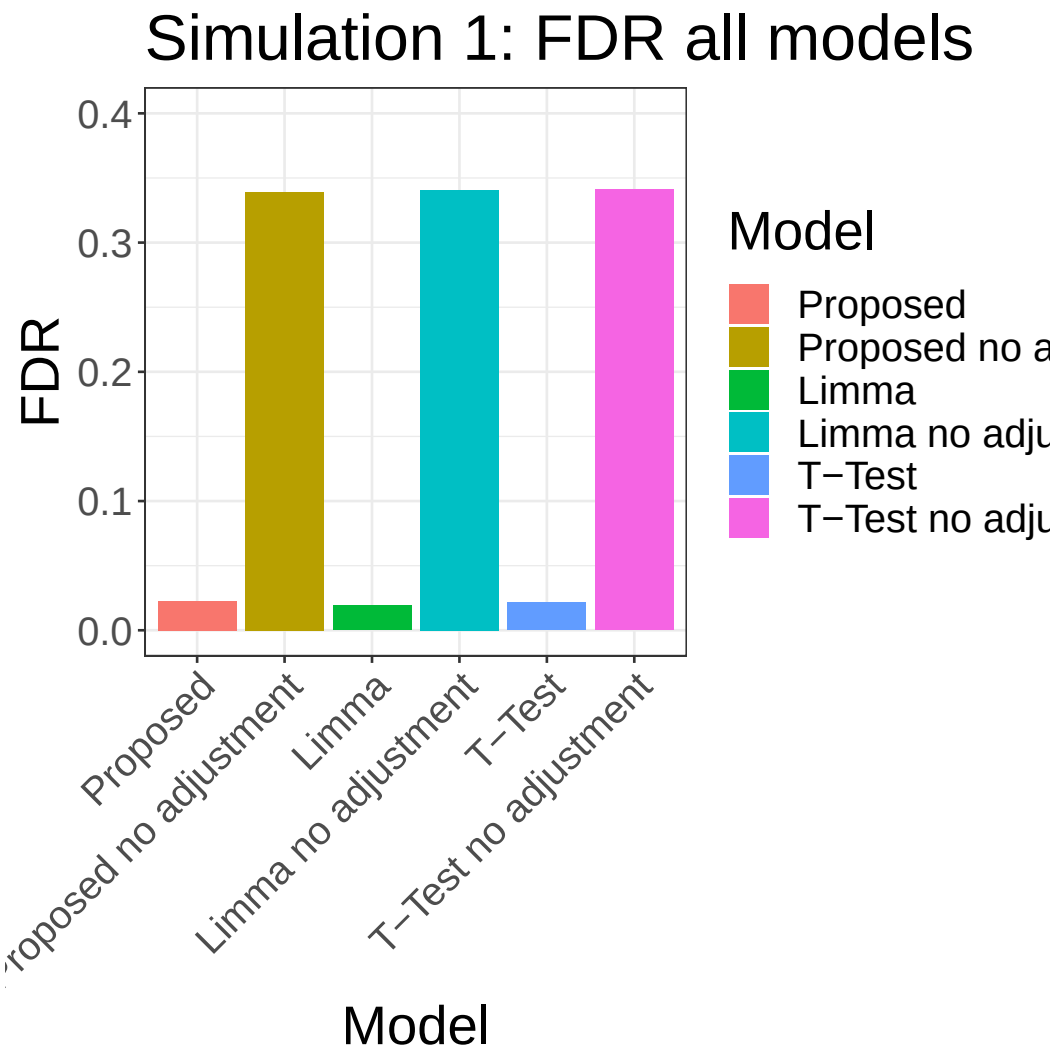
Here we can see a sample PTM with two conditions A and B. The red line represents a same biological replicate. We can simulate different variance's for the conditions.

All charts are made using adjusted pvalue <.05 to designate a significant hit.

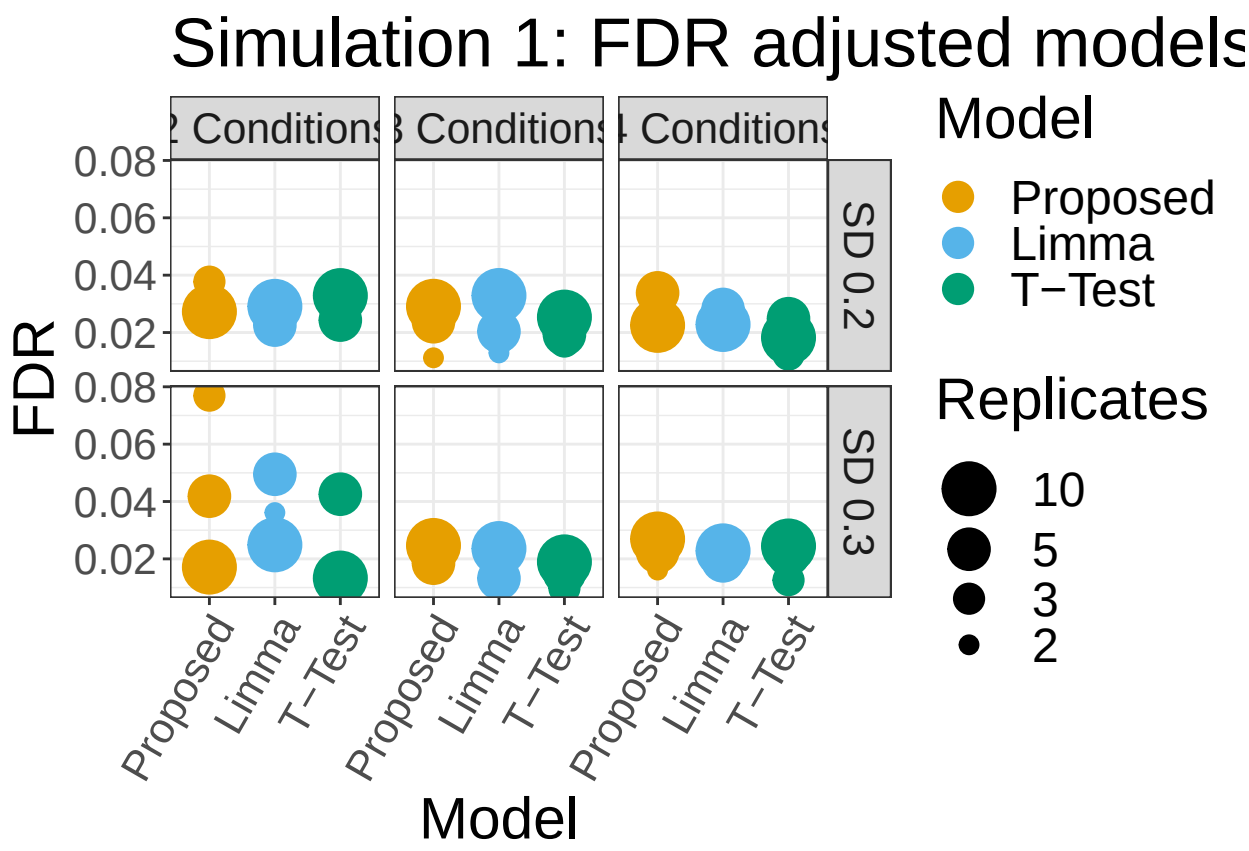
Simulation 1

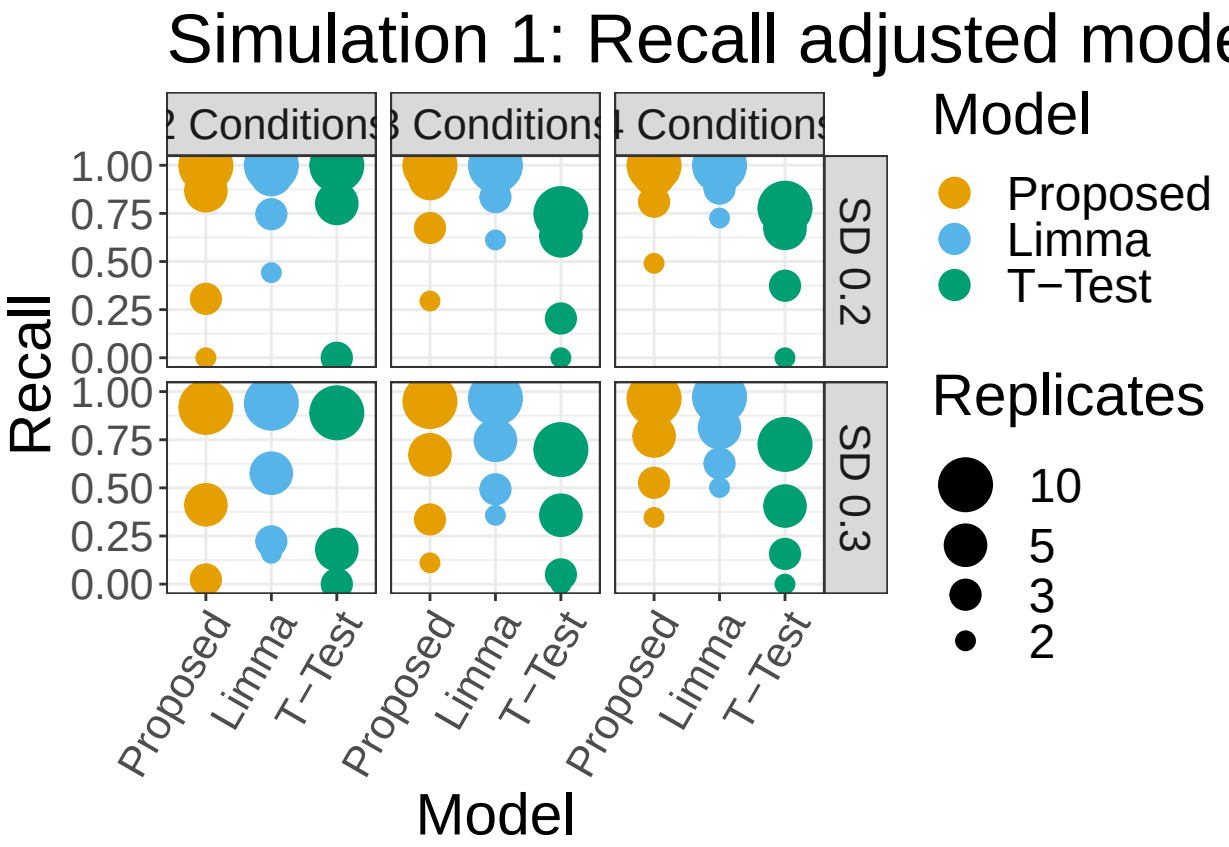
All simulations were ran with half the PTMs being differential, while the other half the difference was due to changes in global protein level. The first simulation is run with the following parameters:

- Mean of log-intensity: 25
- Number of Features: 10 (PTM), 10 (Protein)
- 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: 3, 4, 5
- Number of conditions: 2, 3, 4
- Number of realizations: 500

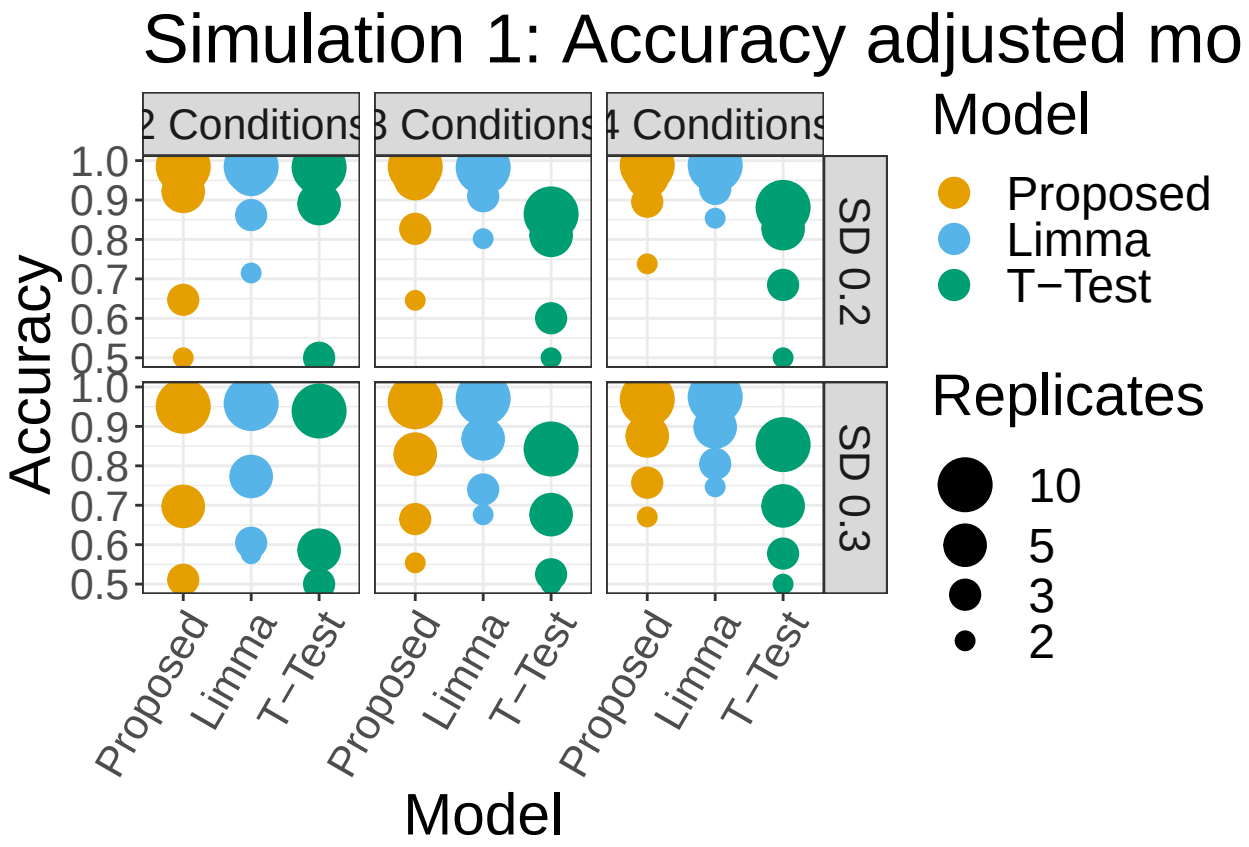


False Discovery Rate (FDR)

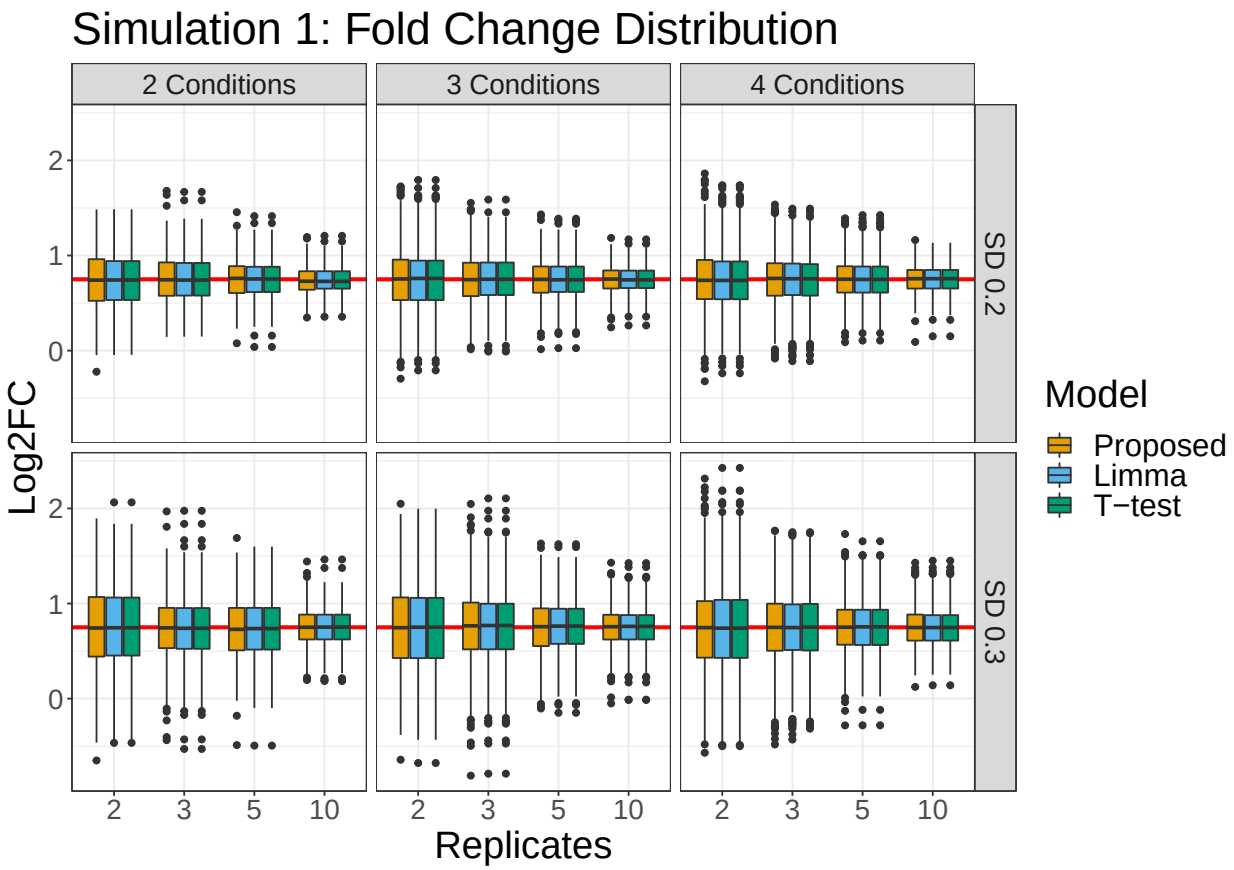

Simulation 1: FDR adjusted models
2 Conditions
3 Conditions
4 Conditions
SD 0.2
SD 0.3
FDR
0.08
0.06
0.04
0.02
Proposed
Limma
T−Test
Model
Model
Proposed
Limma
T−Test
Replicates
10
5
3
2


Simulation 1: Recall adjusted mode
2 Conditions
3 Conditions
4 Conditions
SD 0.2
SD 0.3
Recall
1.00
0.75
0.50
0.25
0.00
Proposed
Limma
T−Test
Model
Model
Proposed
Limma
T−Test
Replicates
10
5
3
2

Recall (TPR)


Simulation 1: Accuracy adjusted mo
2 Conditions
3 Conditions
4 Conditions
SD 0.2
SD 0.3
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
Proposed
Limma
T−Test
Model
Model
Proposed
Limma
T−Test
Replicates
10
5
3
2

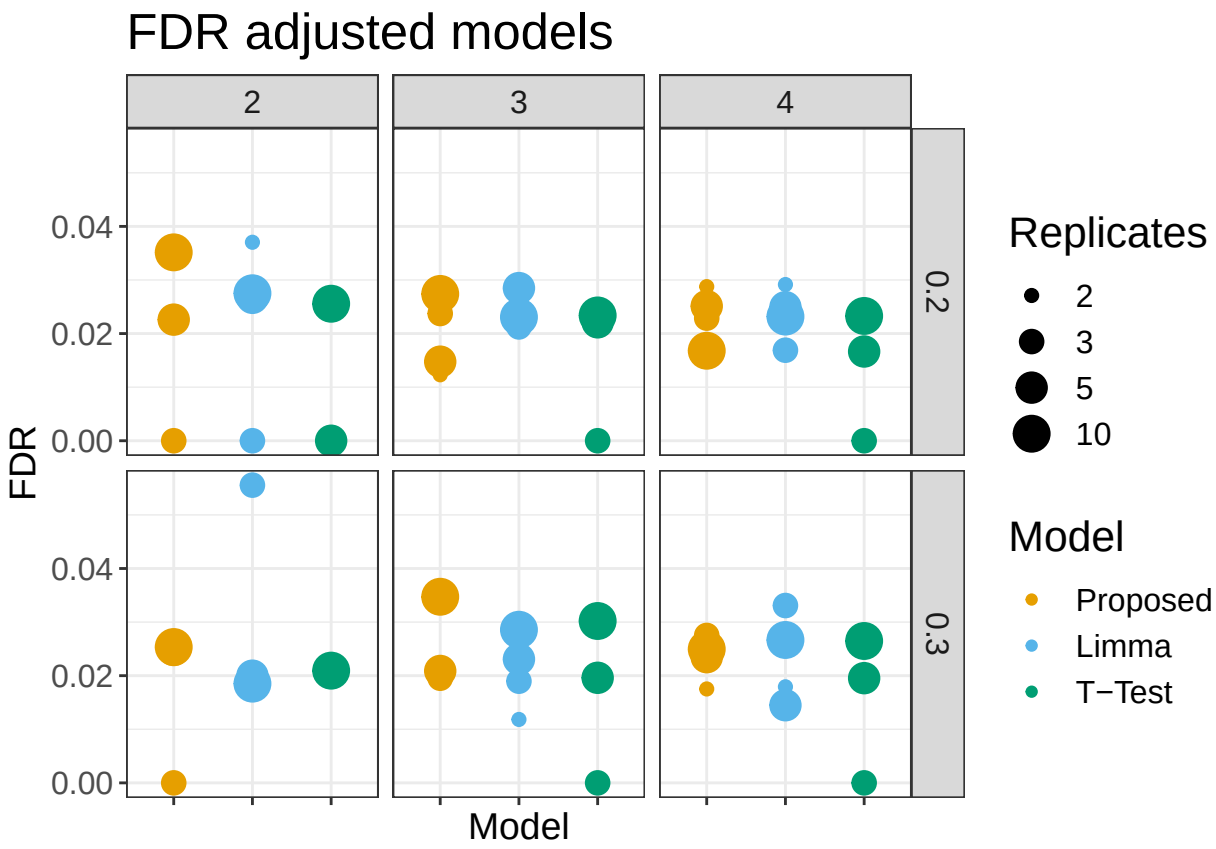
Accuracy



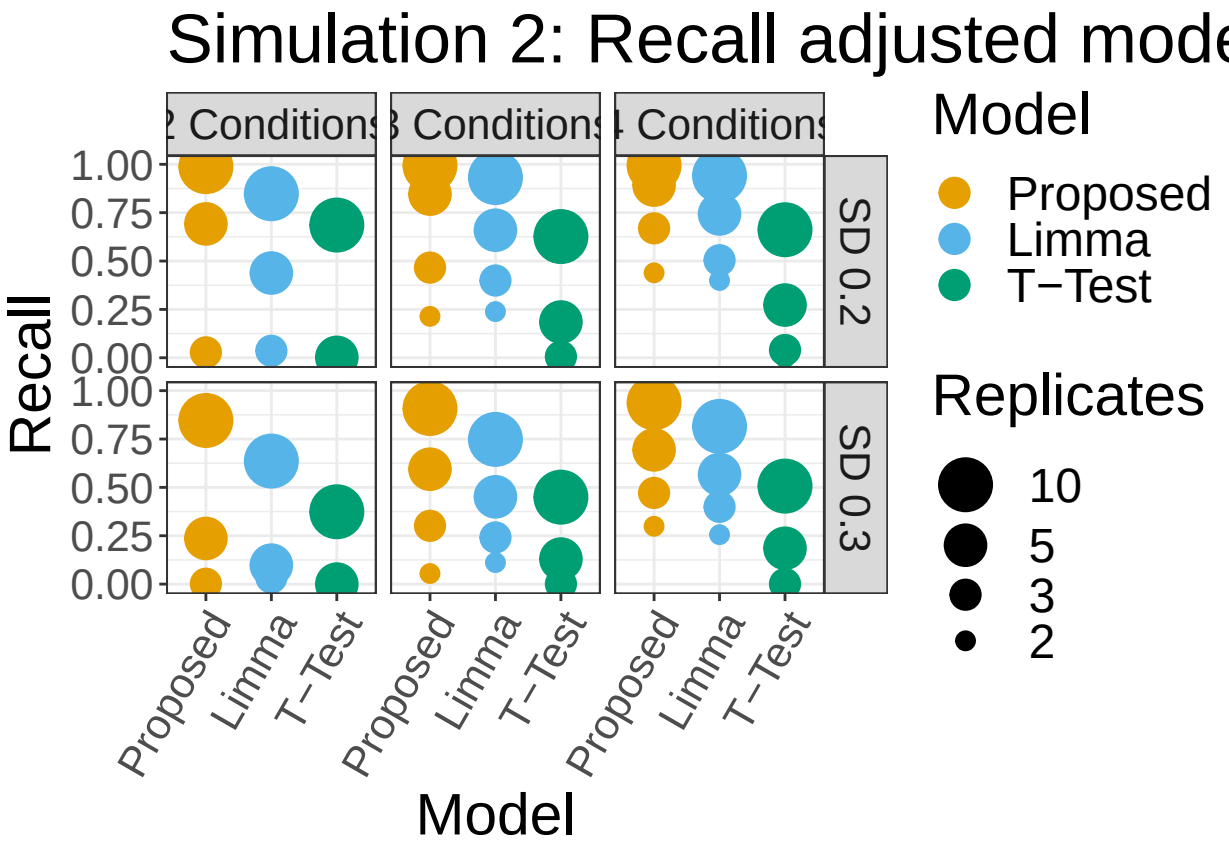
Simulation 2

In this simulation we introduce missing values into the experiment. 20% of the features simulated were selected at random and masked with an NA value. No missing value imputation was used.

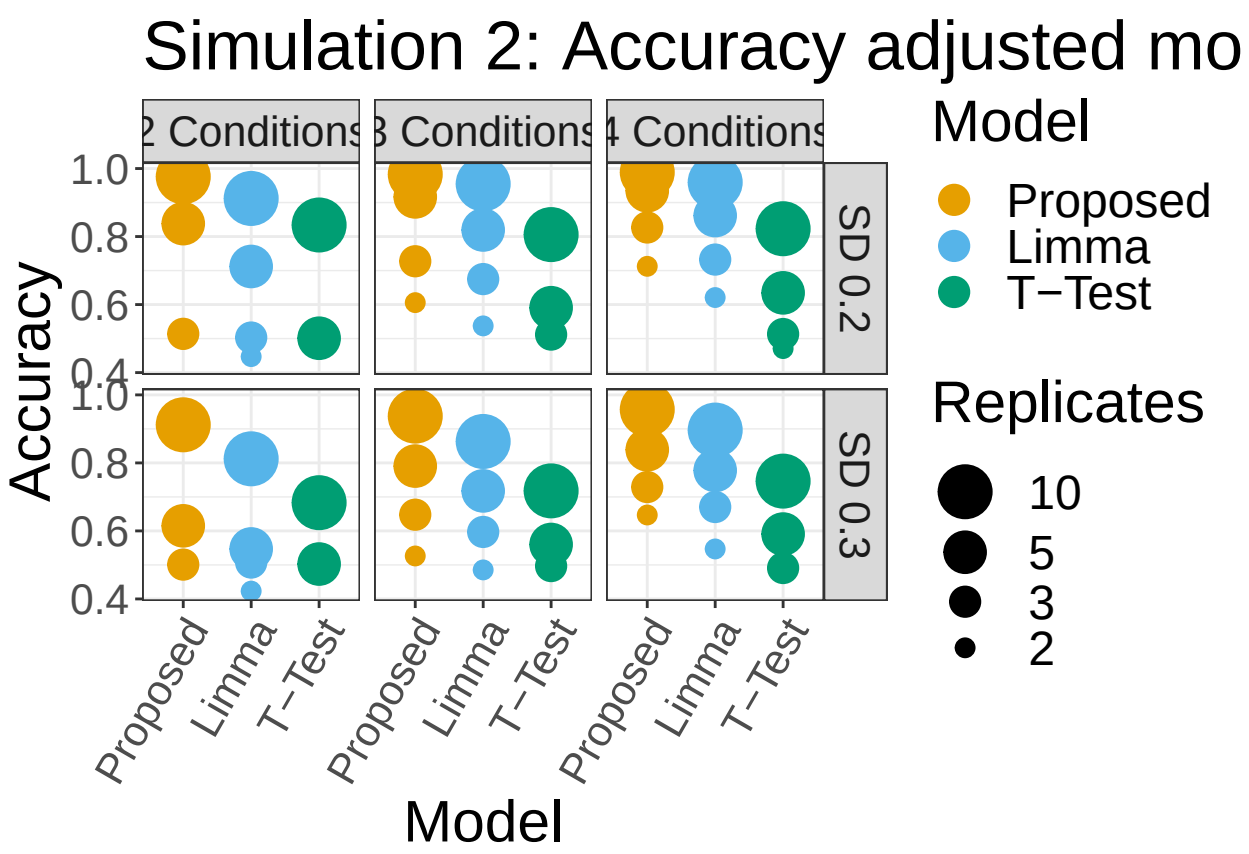
- Mean of log-intensity: 25
- Number of Features: 2 (PTM), 10 (Protein)
- Standard deviations of log-intensities for modified and unmodified peptides: 0.2, 0.3
- Difference in PTM abundance between conditions: 0, 0.75, 1.5, 2.25
- Difference in protein abundance between conditions: 0, 0.75, 1.5, 2.25
- Number of replicates: 2, 4, 6
- Number of conditions: 2, 3, 4
- Number of realizations: 1000
- 10% of features missing at random in PTM and Protein datasets


FDR adjusted models
2
3
4
0.2
0.3
0.04
0.02
0.00
FDR
Model
Replicates
2
3
5
10
Model
Proposed
Limma
T−Test

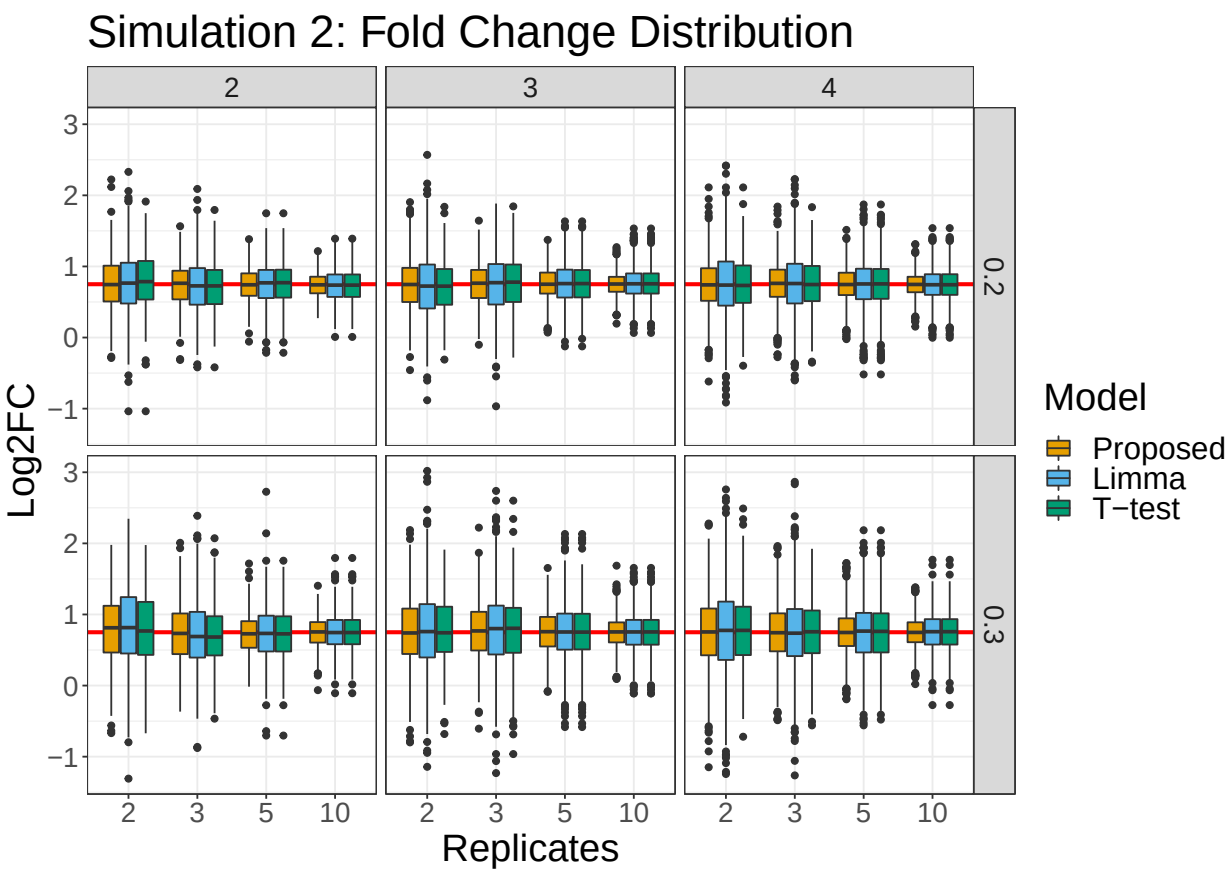
FDR


Simulation 2: Recall adjusted mode
2 Conditions
3 Conditions
4 Conditions
SD 0.2
SD 0.3
Recall
1.00
0.75
0.50
0.25
0.00
Proposed
Limma
T−Test
Model
Model
Proposed
Limma
T−Test
Replicates
10
5
3
2

Recall (TPR)


Simulation 2: Accuracy adjusted mo
2 Conditions
3 Conditions
4 Conditions
SD 0.2
SD 0.3
Accuracy
1.0
0.8
0.6
0.4
Proposed
Limma
T−Test
Model
Model
Proposed
Limma
T−Test
Replicates
10
5
3
2

Accuracy


Simulation 2: Fold Change Distribution
2
3
4
0.2
0.3
Log2FC
3
2
1
0
−1
2
3
5
10
Replicates
Model
Proposed
Limma
T−test