Answers to reviewers

msqrob2TMT: robust linear mixed models for inferring differential abundant proteins in labelled experiments with arbitrarily complex design

Stijn Vandenbulcke Christophe Vanderaa Oliver Crook
Lennart Martens Lieven Clement

Reviewer 3

We thank reviewer 3 for their positive response. There has indeed been a decrease in image resolution during the auto-preparation, but this does not match with the image resolution of the figures we submitted. We will make sure that high resolution images are included in the publication before accepting the final proof.

Reviewer 4

We thank reviewer 4 for their comments. Below, we provide an answer to each concern and indicate how we adapted the manuscript accordingly.

1. Section 2.3. Both FDR and FDP are used in the evaluation. The authors should clarify the differences between these two. In addition, peptide level FDRs and protein level FDRs should be clarified. In Fig. 2A, the dots represent 5% FDR, the authors should clarify why the corresponding FDPs of the dots are so different from each other.

First we would like to point out that discussing peptide-level or protein-level FDR does not seem relevant here since our contribution is not focusing on identification but on inferring differential abundance at the protein level, even when starting from PSM-level data. Indeed, both our PSM-level and protein-level workflows estimate and infer fold changes at the protein-level.

Next, we would like to point out that the FDP is the false discovery proportion i.e., the fraction of false positives in the list of returned DA proteins, which can typically not be evaluated for real experiments as the ground truth is unknown. Therefore, the FDR has been developed, which is the expected FDP i.e., the average of the FDPs that would be obtained when the experiment were to be repeated an infinite number of times. In this contribution we estimate the FDR with the Benjamini Hochberg method, which implies a number of assumptions, e.g. independence

between features (proteins), a known distribution of the test statistic under the null hypothesis and thus assumptions on the distribution of the data, etc.

Interestingly, the FDP can be computed for the spike-in studies that are used to benchmark tools as the spike-in proteins are known to be DA and the background proteins are not DA, which we exploit in our contribution to construct performance curves (sensitivity vs FDP curves). In order to evaluate the FDR control of the different workflows, we would need many repeats of the spike-in study so as to assess if the average FDP equals the nominal FDR-threshold that was used.

As there are no repeats available for the spike-in studies, we have opted to plot the observed FDP at the nominal 5% FDR-level on the performance curves as an indication on how well a workflow can control the FDR.

Note, that the FDP in a specific spike-in study is expected to deviate from the nominal 5% FDR-level by random chance, as the FDR corresponds to the FDP we expect on average rather than the FDP for the single spike-in study.

However, severe deviations of the observed FDP and the nominal FDR level are unexpected and are indicative that a workflow provides poor FDR control as discussed in section 3.2: "These findings are further corroborated by the results at the 5% FDR level, shown in Figure 2 C. All msqrob2 workflows demonstrate high sensitivity and low FDP. Specifically, they recover between 143 - 175 DA hits (true positives,TP) for spike-in UPS proteins across all 6 pairwise comparisons while only reporting between 3 - 9 false positives (FP) for HeLa proteins. As a result, their FDP ranges between 2.1% - 5.1%, suggesting appropriate FDR control at the 5% level. The default msTrawler workflow, however, was only able to recover 142 TP and reported 14 FP, leading to an FDP of 9%. With our refactored import function this improved to 184 TP, 9 FP and an FDP of 4.7%. The summarisation-based workflows DEqMS and MSstatsTMT, reported 169 and 147 TP, respectively, with 39 and 11 FP, resulting in FDPs of 18.8% and 7%, respectively, suggesting improper FDR control by DEqMS."

We clarified the difference between FDP and FDR in our revised manuscript by expanding the end of section 2.3 as follows:

"We also highlight the observed FDP at a 5% FDR threshold. Since the FDR represent the expected FDP, i.e. the average of the FDPs obtained when the spike-in experiment were to be repeated an infinite number of times, an observed FDP that is very far away from 5% is indicative for a workflow that provides poor FDR control."

2. Page 14. "Specifically, they recover between 143 - 175 spike-in UPS proteins as DA." The number of reported proteins is more than the total number of 40 spiked proteins, why? Similarly, are the numbers in Fig. 5c reported proteins.

The number of true positives and false positives (for both spike-in datasets) are reported across all pairwise comparisons. Hence, the maximum number of differentially abundant proteins is 40 x 6 = 240.

The above sentence now reads as follows in the revised version of our manuscript: "Specifically, they recover between 143 - 175 DA hits (true positives,TP) for spike-in UPS proteins across all 6 pairwise comparisons. . . "

3. Fig. 3A. Why are yeast proteins identified from the MSstatsTMT spike-in data?

We thank the reviewer for spotting this mistake. We have adapted the plot title and the plot caption accordingly, which now reads as "HeLa background proteins" (plot title panel 3A) and "Figure 3: Boxplots showing the log2 FC distributions for the spike-in and non-spike-in proteins, focusing on two comparisons with the highest and lowest difference in spike-in concentrations. Fold changes are estimated by DEqMS, msqrob2TMT, MSstatsTMT, and msTrawler workflows. The grey dotted line is the true log2 FC for the comparison. Panel A: non spike-in proteins (HeLa); Panel B: spike-in proteins (UPS)." (plot caption).

4. Page 3. No references are provided for DEqMS and msTrawler when they are first mentioned.

We have moved the citation to the first occurrence in the text in the revised version of our manuscript.

5. Page 4, section 2.1.1. "A reference sample was prepared by combining the diluted UPS1 peptide samples with 50 μg of SILAC HeLa peptides." The concentration of the diluted UPS1 peptide sample is not given.

We clarified that the reference sample contained 286.5 fmol of the combined UPS samples (ie the average of the 4 spike-in amounts). We also fixed the inconsistencies between peptide amount and peptide concentration.

The paragraph now reads as follows: "The spike-in dataset was obtained from MassIVE (RMSV000000265) and has the following design: 500, 333, 250, and 62.5 fmol of UPS1 peptides were spiked into 50 μg of SILAC HeLa peptides. This series forms a dilution gradient of 1, 0.667, 0.5, and 0.125 relative to the highest UPS1 peptide amount (500 fmol). A reference sample was prepared by combining the diluted UPS1 peptide samples (286.5 fmol) with 50 μg of SILAC HeLa peptides."

6. Page 4, section 2.1.1. Details of database search methods, such as software version, parameter settings, should be provided.

We added the following sentence "The MS data were searched by the authors using Proteome Discoverer 2.2.0.388 (Thermo Fisher Scientific) and Mascot Server 2.6.1 (Matrix Science, London, UK)." However, more information has not been provided in the original publication from which we retrieved the searched data.

7. Page 5. Section 2.1.2. "For this dataset the ground truth is know: a true positive is a yeast protein that is returned as significant while a significant mouse protein is a false positive." The sentence is confusing.

We adjusted this phrase in the revised version of our manuscript, which now reads as follows: "For this dataset the ground truth is known: yeast proteins were spiked in the samples in different amounts, while mouse proteins were added as a constant background. Hence, any yeast protein that is returned as significant is a true positive, and any mouse protein that is returned as significant is a false positive."

8. Page 6. "The log2-transformed PSM intensities are first centered by subtracting the corresponding median of the log2 PSM intensities from their corresponding spectrum." It is unclear what are their corresponding spectra.

We adjusted this phrase in the revised version of our manuscript as follows: "This approach can be regarded as a one-step median polish that corrects for PSM-specific effects (spectra effects). The PSM-specific effect is shared across all TMT intensities from the same spectrum, which can be estimated by their corresponding median. Hence, the median sweep algorithm corrects TMT intensities for PSM-specific effects by subtracting their corresponding median."

9. Fig. 2. The resolution of the figures are low.

We agree the PDF generated by the editor contains low-quality figures, but this does not match with the image resolution of the figures we submitted. We will make sure that high resolution images are included in the publication before accepting the final proof.