Received: 15 November 2021 | Revised: 23 January 2022 | Accepted: 27 January 2022

DOI: 10.1002/jcb.30225

TOOLBOX

Journal of Cellular Biochemistry WILEY

PBLMM: Peptide-based linear mixed models for differential expression analysis of shotgun proteomics data

Kevin Klann | Christian Münch

Institute of Biochemistry II, Faculty of Medicine, Goethe University Frankfurt, Frankfurt am Main, Germany

Correspondence
Christian Münch, Institute of Biochemistry II, Faculty of Medicine, Goethe University Frankfurt, Theodor-Stern-Kai 7/Bldg 75, 60590 Frankfurt am Main, Germany.
Email: ch.muench@em.uni-frankfurt.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: 259130777-SFB1177, 390339347-Emmy Noether Programme, 403765277;
Seventh Framework Programme, Grant/Award Number: ERC StG 803565

Abstract
Here, we present a peptide-based linear mixed models tool—PBLMM, a standalone desktop application for differential expression analysis of proteomics data. We also provide a Python package that allows streamlined data analysis workflows implementing the PBLMM algorithm. PBLMM is easy to use without scripting experience and calculates differential expression by peptide-based linear mixed regression models. We show that peptide-based models outperform classical methods of statistical inference of differentially expressed proteins. In addition, PBLMM exhibits superior statistical power in situations of low effect size and/or low sample size. Taken together our tool provides an easy-to-use, high-statistical-power method to infer differentially expressed proteins from proteomics data.

KEYWORDS
bioinformatics, data analysis, differential expression, proteomics, statistics

1 | INTRODUCTION

The advances in mass spectrometry instrumentation nowadays allow for the quantification of multiple peptides per protein (up to a few hundred) during shotgun proteomics experiments. Quantification accuracy of different peptides during mass spectrometry runs is highly dependent on the physical properties of individual peptides and might differ from run to run due to technical constraints. Therefore, outlier peptides can bias the quantification accuracy of the whole protein and need to be carefully assessed during differential expression analysis. However, in most downstream analysis workflows for differential expression analysis, the peptide level information is ignored and summed to a single protein quantification used for statistical analysis.[1,2] While these workflows are commonly applied and represent easy and intuitive methods for statistical analysis, they discard the obtained information on peptide level, which leads to loss of statistical power.[3] In recent years, the first approaches using linear models for differential expression analysis have been transferred from microarray experiments to proteomics.[4–7] These are capable of carrying out statistical analyses on the peptide level. However, only a few packages use some of the peptide information for statistical analysis[8,9] and are often only available for certain workflows, such as label-free quantification.[9] Linear mixed models (LMMs) in general have been suggested and used for statistical inference before (notably, the exact model differs from study to study) and despite them having shown great statistical power,[3,6,8,9] there is no easy-to-use package or standalone application published so far that is also applicable to multiplexed proteomics. To solve this issue we present peptide-based linear mixed models (PBLMM), a standalone desktop application for differential protein expression analysis from shotgun proteomics experiments, especially those applying isobaric labellings, such as tandem mass tags (TMT) or iTRAQ. We compare PBLMM to currently used tools, such as MSstatsTMT[10] or limma[4] and found PBLMM to provide statistical benefits over these methods, depending on the use-case.

1.1 | Statistical model of PBLMM

In the implemented statistical model, the expression of each protein is separately modelled by a linear mixed-effects model:

$$y_{i,j} = \beta_0 + \beta X + u_i + \varepsilon_{i,j},$$

where $y_{i,j}$ denotes the jth measurement of expression of peptide i, β_0 is the individual protein's global intercept, βX is the linear combination of indicator variables encoding categorical experimental conditions, u_i is the additive random intercept of peptide i with $u_i \sim N\,(0, \sigma^2_{\text{Peptide}})$, and $\varepsilon_{i,j}$ are residual errors with $\varepsilon_{i,j} \sim N\,(0, \sigma^2_\varepsilon)$. Note that this collapses to ordinary linear regression when there are no multiple peptide measurements per protein.

In the absence of further experimental conditions, the variance of the response variable $y_{i,j}$ can be described by the sum of the variance components (σ^2) peptide, technical replicate (TechRep), and multiplex, as well as unexplained residual variance σ^2_ε:

$$\sigma^2_{y_i} = \sigma^2_{\text{Peptide}} + \sigma^2_{\text{TechRep}} + \sigma^2_{\text{Multiplex}} + \sigma^2_\varepsilon.$$

The variance components reduce when no technical replicates and/or different multiplexes are present. This setup makes PBLMM aware of most experimental designs that are commonly used in proteomics.

Since the input matrix is $\log_2$ transformed, the treatment coefficients from the models can be interpreted as $\log_2$ fold changes and p values for the main treatment effects can be extracted. The null hypothesis tested is that the coefficient for the tested term equals zero and the factor is not meant to explain the protein expression. Therefore a low p value indicates the importance of the factor and a low likelihood of the fold change is 0 for the tested treatment. To control the false discovery rate (FDR), we applied multiple testing corrections by the Benjamini–Hochberg FDR method.[11] The application automatically calculates differential analysis between all condition pairs possible.

2 | RESULTS

We created two implementations of the PBLMM statistical model described above (Figure 1A): (i) for user-friendly analysis, we implemented the PBLMM algorithm into a standalone desktop application with a graphical user interface and (ii) a Python package containing additional parameters, processing steps and pipelining features to facilitate customisable data analysis workflows.

To test how PBLMM performs compared to other commonly used methods, we created a ground truth dataset, consisting of a TMTpro 12-plex containing spiked-in *Escherichia coli* digests in a human background in different known ratios (Figure 1B). We then performed differential expression analysis using different classical protein level statistics and PBLMM: (i) sum-based protein rollup (peptides are summed for each

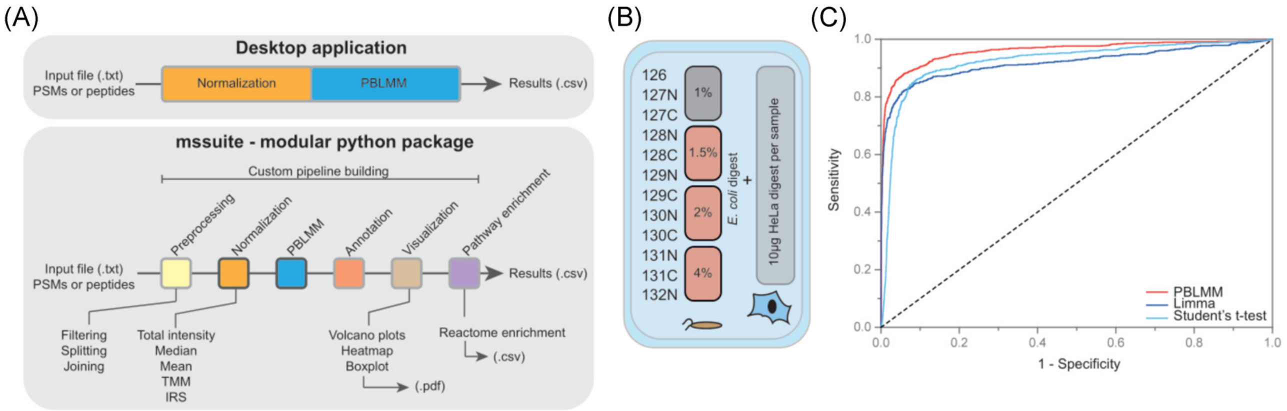

FIGURE 1 Statistical evaluation of peptide-based linear mixed model (PBLMM). (A) Scheme of PBLMM desktop application and modular python package. (B) Scheme of the ground truth dataset composition. Different amounts of *Escherichia coli* digest were spiked into a fixed HeLa background. Peptides were tandem mass tag labelled and fractionated by offline reverse phase HPLC into 24 fractions. (C) Receiver operating characteristic curves of p values generated by different statistical tools/tests as predictors of species. Statistics for samples with twofold changes were used for this graph

protein before statistical inference), followed by Student's *t*-test, (ii) sum-based protein rollup, followed by limma,[4] and (iii) PBLMM. Receiver operating characteristic (ROC) curve analysis showed that PBLMM outperforms the alternative classic methods on our test dataset (Figure 1B). Notably, we found that the difference between statistical approaches was less pronounced with very high effect sizes (e.g., fourfold). We thus performed our analysis with lower effect sizes throughout this study.

2.1 | PBLMM shows advantages with small effect size and low replicate number

While methods using protein-based statistics exhibit high statistical power in situations where enough replicates are present to sufficiently estimate variances of indicator variables, they inherently lack statistical power in experimental designs that rely on low numbers of replicates or study low fold changes across conditions. In these cases, the use of peptide-level data provides additional statistical power. For each sample and protein, multiple peptides are measured giving a better complete view of the different variance sources, biological, or technical. We tested PBLMM against the current state-of-the-art tool MSstatsTMT[5,10] (Figure 2A). MSstatsTMT uses flexible LMMs to infer differential expression between biological conditions from TMT data. However, the LMMs are fitted on protein levels, previously inferred from an additive linear model on peptide level data. Here we used our fractionated ground truth dataset from before and validated the statistical power of both tools. In the ROC analysis, both tools show comparable performance. However, when we applied standard FDR cut-offs, such as 0.05 or 0.01 (either alone or in combination with additional fold change cut-offs), protein-level statistics failed to detect any significantly changed proteins at a small effect size of 1.5-fold (Figure 2B). This effect has been already discussed by others and represents a common problem in proteomics experiments.[12] In contrast, the enhanced statistical power of our peptide-based model was able to detect several hundred significantly changed proteins. While both methods performed comparably at a higher effect size with three replicates, we also observed more pronounced differences during the analysis of only two replicates (Figure 2C). Here, only the peptide-based model was able to correctly detect differentially expressed proteins, although with a slightly inflated empirical FDR that could be easily controlled by applying additional fold change cut-offs. Since large experiments with tens to hundreds of conditions are becoming more and more popular and naturally suffer from lower replicate numbers,[13,14] statistical models for their analyses become increasingly important.

We next hypothesised that also the estimation of technical variation between runs can benefit from peptide data. Thus, we generated another ground truth dataset and measured the resulting multiplex three times as technical replicates (Figure 3A). While the differentially expressed proteins clearly separated from the background (i.e., human proteins) when analysing differences inside one multiplex (Figure 3B), the data points started to converge when looking at the data across the different replicate injections (Figure 3C). The technical variation of the background proteins between MS runs was higher than the effect size, thus masking the real effects. Linear models are able to divide these effects and calculate the technical and biological variance separately, leading to high statistical power. When we applied our peptide-based model in conjunction with internal reference scaling preprocessing,[15] we observed an improvement in statistical power compared to the protein level quantification, although nine replicates (three replicates for each three MS runs) were present for each condition (Figure 3D).

Taken together, we found that PBLMM performed comparably to other state-of-the-art statistical tools and provides distinct advantages in situations commonly observed in biological or medical experimental designs, such as experiments with low effect sizes or low numbers of replicates. These advantages are beneficial for large scale experiments, which would not be feasible with the currently required high number of replicates, and conditions with small effects sizes.

3 | DISCUSSION

Statistical data analysis of proteomics experiments needs adjustments for each experiment since it is heavily influenced by multiple parameters. The effect size is not only determined by the biological effects but also strongly affected by instrumentation. Advances in instrumentation are occurring rapidly and many different instruments are used in proteomics facilities worldwide, each influencing the effect size and technical variance in their own way. In addition, applied measurement methods, offline fractionation, and sample preparation are varying a lot across laboratories. Therefore, each experiment may have its own statistical needs that may not be covered by single tools. Consequently, we think that the choice of the statistical tools should ideally depend on the experimental setup.

LMMs have proven to be able to account for several of these aspects, like technical or subject variance, and therefore provide more advanced hypothesis testing. We showed that PBLMM is able to perform similarly or better than state-of-the-art methods like MSstatsTMT

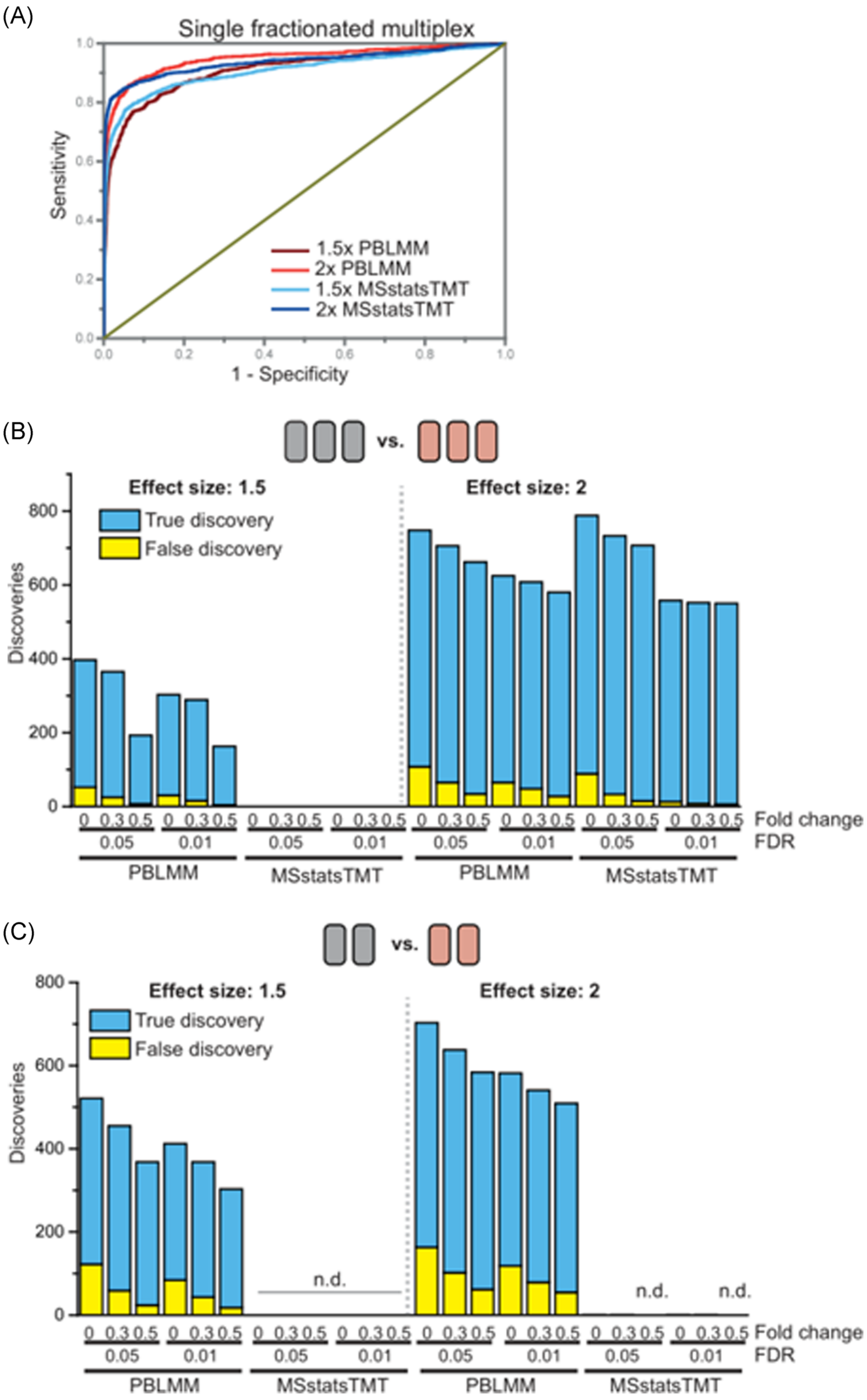

FIGURE 2 Peptide-based linear mixed model (PBLMM) displays advantages with small effects sizes and low replicate numbers. (A) Receiver operating characteristic analysis of statistical parameters of PBLMM compared to MSstatsTMT with two effect sizes: 1.5-fold effect size and 2-fold effect size. (B and C) Number of true and false discoveries with different commonly used false discovery rates (FDR) and fold change cut-offs. Grey and red rectangles schematically represent number of replicates used in statistical analysis: (B) $n = 3$; (C) $n = 22$. Blue: true discoveries; Yellow: false discoveries

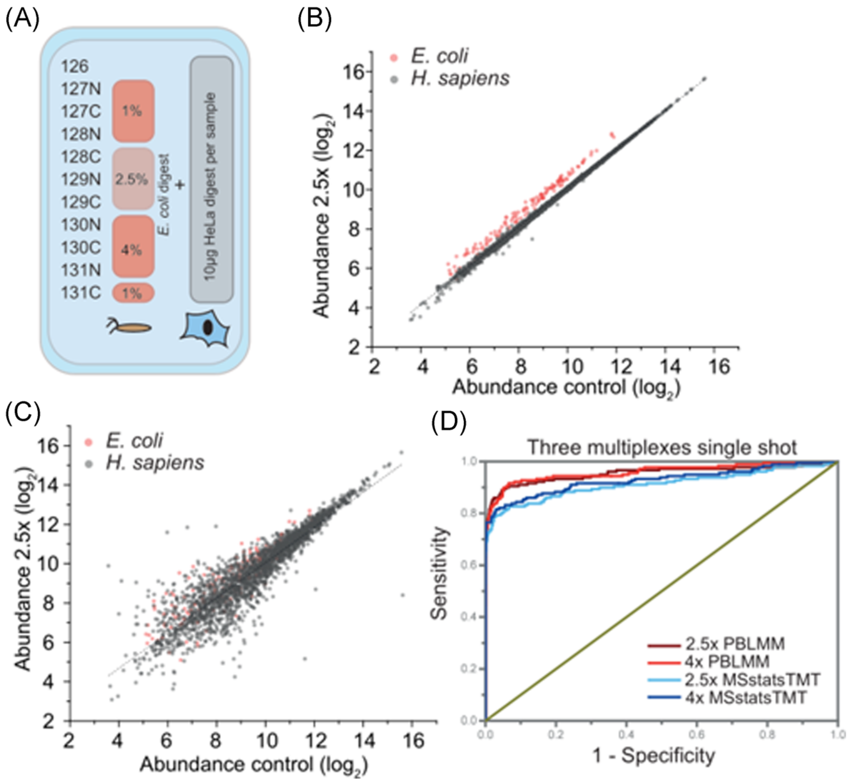

FIGURE 3 Peptide data allows accurate estimation of technical variance. (A) Scheme of ground truth dataset composition. *Escherichia coli* peptides were spiked into a fixed HeLa background, tandem mass tag (TMT) labelled and measured in three technical replicates. (B) Comparing abundances of 1% (control) and 2.5% (2.5×) samples in one multiplex. The differentially expressed *E. coli* proteins are shown in red. (C) Comparing abundances of 1% (control) and 2.5% (2.5×) samples from two different technical replicates. (D) Receiver operating characteristic analysis of *p* values generated by PBLMM and MSstatsTMT

depending on the experimental setup. Strikingly, we found that, for several scenarios, using the peptide information directly for differential expression analysis, strongly enhanced the statistical power, while maintaining a low FDR. Especially testing biological conditions with low effect sizes (as observed in mild treatments) or a low number of replicates (as observed in high-throughput experiments) benefited from the additional information provided by the individual peptides. We anticipate that the power of statistical tools might not be adequately reflected solely by ground truth datasets, however, these datasets represent the only quantifiable source of statistical power and accuracy together with purely simulated datasets. Importantly, PBLMM is easily accessible as an interactive desktop tool and thus allows it to be used broadly used in different analyses pipelines and for all different types of input data.

ACKNOWLEDGEMENTS

The authors thank S. Koschade for the discussion and input. Christian Münch acknowledges support from the European Research Council under the European Union's Seventh Framework Programme (ERC StG 803565), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project ID 390339347 (Emmy Noether Program), project ID 259130777 (CRC1177 Selective Autophagy) and project ID 403765277 (mass spectrometer).

CONFLICT OF INTERESTS

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The source code and implementations are made freely accessible via Github under https://github.com/klannk/mssuite and https://github.com/klannk/Peptide_based_LMM. All proteomics data is will be shared upon request.

ORCID

Kevin Klann http://orcid.org/0000-0003-2276-8128
Christian Münch https://orcid.org/0000-0003-3832-090X

REFERENCES

1. Mueller LN, Brusniak M-Y, Mani DR, Aebersold R. An assessment of software solutions for the analysis of mass spectrometry-based quantitative proteomics data. *J Proteome Res*. 2008;7(1):51-61. doi:10.1021/pr700758r
2. Liu H, Sadygov RG, Yates JR. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem*. 2004;76(14):4193-4201. doi:10.1021/ac0498563
3. Goeminne LJE, Argentini A, Martens L, Clement L. Summarization vs peptide-based models in label-free quantitative proteomics: performance, pitfalls, and data analysis guidelines. *J Proteome Res*. 2015;14(6):2457-2465. doi:10.1021/pr501223t
4. Ritchie ME, Phipson B, Wu D, et al. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. doi:10.1093/nar/gkv007

5. Choi M, Chang C-Y, Clough T, et al. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*. 2014;30(17):2524-2526. doi:10.1093/bioinformatics/btu305
6. D'Angelo G, Chaerkady R, Yu W, et al. Statistical models for the analysis of isobaric tags multiplexed quantitative proteomics. *J Proteome Res*. 2017;16(9):3124-3136. doi:10.1021/acs.jproteome.6b01050
7. Daly DS, Anderson KK, Panisko EA, et al. Mixed-effects statistical model for comparative LC−MS proteomics studies. *J Proteome Res*. 2008;7(3):1209-1217. doi:10.1021/pr070441i
8. Zhu Y, Orre LM, Tran YZ, et al. DEqMS: a method for accurate variance estimation in differential protein expression analysis. *Mol Cell Proteomics*. 2020;19(6):1047-1057. doi:10.1074/mcp.TIR119.001646
9. Goeminne LJE, Sticker A, Martens L, Gevaert K, Clement L. MSqRob takes the missing hurdle: uniting intensity- and count-based proteomics. *Anal Chem*. 2020;92(9):6278-6287. doi:10.1021/acs.analchem.9b04375
10. Huang T, Choi M, Tzouros M, et al. MSstatsTMT: statistical detection of differentially abundant proteins in experiments with isobaric labeling and multiple mixtures. *Mol Cell Proteomics*. 2020;19(10):1706-1723. doi:10.1074/mcp.RA120.002105
11. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*. 1995;57(1):289-300.
12. Pascovici D, Handler DCL, Wu JX, Haynes PA. Multiple testing corrections in quantitative proteomics: a useful but blunt tool. *Proteomics*. 2016;16(18):2448-2453. doi:10.1002/pmic.201600044
13. Nusinow DP, Szpyt J, Ghandi M, et al. Quantitative proteomics of the cancer cell line encyclopedia. *Cell*. 2020;180(2):387-402.e16. doi:10.1016/j.cell.2019.12.023
14. Johnson ECB, Dammer EB, Duong DM, et al. Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nat Med*. 2020;26(5):769-780. doi:10.1038/s41591-020-0815-6
15. Plubell DL, Wilmarth PA, Zhao Y, et al. Extended multiplexing of tandem mass tags (TMT) labeling reveals age and high fat diet specific proteome changes in mouse epididymal adipose tissue. *Mol Cell Proteomics*. 2017;16(5):873-890. doi:10.1074/mcp.M116.065524

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Klann K, Münch C. PBLMM: Peptide-based linear mixed models for differential expression analysis of shotgun proteomics data. *J Cell Biochem*. 2022;123:691-696. doi:10.1002/jcb.30225