bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Utilizing `scp` for the analysis and replication of single-cell proteomics data

Christophe Vanderaa[1] and Laurent Gatto[1]

[1]Computational Biology and Bioinformatics Unit (CBIO), de Duve Institute, UCLouvain, Belgium

Email: laurent.gatto@uclouvain.be

Abstract

Introduction Mass spectrometry-based proteomics is actively embracing quantitative, single cell-level analyses. Indeed, recent advances in sample preparation and mass spectrometry (MS) have enabled the emergence of quantitative MS-based single-cell proteomics (SCP). While exciting and promising, SCP still has many rough edges. The current analysis workflows are custom and build from scratch. The field is therefore craving for standardized software that promotes principled and reproducible SCP data analyses.

Areas covered This special report represents the first step toward the formalization of standard SCP data analysis. `scp`, the software that accompanies this work can successfully reproduces one of the landmark data in the field of SCP. We created a repository containing the reproduction workflow with comprehensive documentation in order to favor further dissemination and improvement of SCP data analyses.

Expert opinion Reproducing SCP data analyses uncovers important challenges in SCP data analysis. We describe two such challenges in detail: batch correction and data missingness. We provide the current state-of-the-art and illustrate the associated limitations. We also highlights the intimate dependence that exists between batch effects and data missingness and provides future tracks for dealing with these exciting challenges.

Keywords: mass spectrometry, proteomics, single-cell, batch correction, imputation, R, Bioconductor, software, reproducible research.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 Article highlights

- Single-cell proteomics (SCP) is emerging thanks to several recent technological advances, but further progress is lagging due to principled and systematic data analysis.
- This work offers a standardized solution for the processing of SCP data demonstrated by the reproduction of a landmark SCP work.
- Two important challenges remain: batch effects and data missingness. Furthermore, these challenges are not independent and therefore need to be modeled simultaneously.

2 Introduction

High-throughput single-cell assays are instrumental in highlighting the biology of heterogeneous cell populations, tissues and cell differentiation processes. Single cell RNA sequencing (scRNA-seq) is a prominent player, thanks to is throughput, technical diversity, and computational tools that support its analysis and interpretation. scRNA-seq is however blind to the many biologically active gene products, proteins and their many proteoforms. Mass spectrometry-based approaches to study the proteome of single cells are emerging (Slavov, 2021, 2020; Kelly, 2020; Ctortecka and Mechtler, 2021), using the wide range of possibilities offered by the technology, including miniaturized sample preparation, labeled and label-free quantitation, as well as data dependent and independent approaches. All these avenues promise to be valuable contributions to the single cell tool kit.

In this work, we will focus on the processing of mass spectrometry-based single cell quantitative data, as produced from the raw data using widely used tools such as, for example, MaxQuant (Tyanova et al., 2016) or Proteome Discoverer (Thermo Fisher Scientific). As expected for a young and fast evolving field such as single-cell proteomics (SCP), there are yet no best practice nor any consensus as to how to adequately process such data. Some studies started from protein and peptide tables as produced by MaxQuant followed by manual data manipulation using Excel (Zhu et al., 2019; Cong et al., 2020), others proceed with Perseus (Zhu et al., 2018b,a; Brunner et al., 2020), other use private in-house scripts (Dou et al., 2019; Zhu et al., 2019), while others publish they custom scripts openly (Schoof et al., 2019; Specht et al., 2021). In this work, will present the reproduction of the open-source scripts of SCoPE2 published by Specht et al. (2021) and their implementation as a formal R/Bioconductor package named `scp`. Reproducing this work allows the formalization and standardization of the current SCP data processing pipeline, but it also brings to light two

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

important challenges for SCP data analysis that we will address in the Expert Opinion section.

3 Reproducing the SCoPE2 analysis

We focused on reproducing the SCoPE2 analysis provided in Specht et al. (2021) since this work puts a milestone in the SCP field by reporting the acquisition of over a thousand single-cells and proving that SCP has reached its potential of becoming a high-throughput technology (Specht and Slavov, 2018). The authors openly shared their raw and quantitative data, as well as their processing scripts. Furthermore, they implemented new metrics and quality controls that could broadly benefit to the field. Although the provided code could fully repeat their results, it is difficult to read for non expert programmers and lacks modularity making it tedious to reuse and hard to adapt and extend. We therefore decided to provide a standardized and modularized framework to reproduce this analysis and hence offer a common ground for SCP data analysis and method development. Our data structure is relying on two curated R/Bioconductor (Huber et al. (2015)) data classes: `QFeatures` (Gatto (2020)) and `SingleCellExperiment` (Amezquita et al. (2019)). `QFeatures` is a data object model dedicated to the manipulation and processing of MS-based quantitative data. It explicitly records the successive steps to allow users to navigate up and down the different MS levels. `SingleCellExperiment` is another data object model designed as an efficient data container that serves as an interface to state-of-the-art methods and algorithms for single-cell data. Our framework combines the two classes to inherit from their respective advantages. Based on this data framework, we built two pieces of software: `scpdata` and `scp`.

The `scp` package extends the functionality of `QFeatures` to SCP applications. For instance, it includes functionality that was implemented in SCoPE2, such as normalization by a reference channel, filtering single-cells based on the median coefficient of variation, or filtering of peptide-spectrum matches (PSM) based on the single-cell to carrier ratio (SCR). A core feature of the `scp` package is the conversion of standard data tables, like those exported by MaxQuant or ProteomeDiscover (Thermo Fisher Scientific), to `scp` formatted data objects along with sample metadata. `scpdata` disseminates SCP data sets formatted using our data structure. The purpose of `scpdata` is three-fold. First, it is an ideal platform for data sharing and hence lays the ground for open and reproducible science in SCP. For instance, the package provides, among others, the PSM, peptide and protein data supplied in Specht et al. (2021) that was used for this replication study. Second, it facilitates the access for developers to SCP data to build and benchmark new methodologies. Finally, the `scpdata`

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

package facilitates the access for new users to data in the context of training and demonstration.

The first step of the reproduction was to retrieve the SCoPE2 data. The data are hosted on Google Drive and are clearly linked from the authors' web page (https://scope2.slavovlab.net/docs/data). We formatted the data set using `scp` and included it in `scpdata` along with comprehensive documentation about data content, data acquisition and data collection. This is true for any data set in `scpdata`. Next, we retrieved the SCoPE2 code from the authors' GitHub repository[1] and formalized the key steps of the workflow (Figure 1A). Most steps implemented in SCoPE2 are routinely performed in bulk proteomics and are easily handled by existing software such as `QFeatures`. The reuse of existing code is essential in software development because it allows the developer to focus on the innovative aspects of its research field without losing time reinventing the wheel (Huber et al. (2015)). Next, we implemented the few missing steps in `scp` and provided clear documentation and examples. Finally, we wrote a new workflow that fully reproduced the results of SCoPE2 script, using our standardized software. The output obtained after running the `scp` workflow leads to very similar results compared to the data provided by the authors (Figure 1B). The set of filtered cells and proteins are almost identical. The final processed data using the two workflows shows high similarity with most differences close to zero. A small proportion of the processed protein expression values show important differences between the two workflows. Since this is not observed at the peptide expression level, we suspect those large difference are the consequence of unstable data imputation and/or batch correction. A detailed report about the reproduction of the SCoPE2 analysis using `scp` can be found on GitHub[2](Vanderaa and Gatto, 2021). This report includes the code used and some comprehensive documentation to give the reader a good understanding of the underlying processes. It also gives additional comments on each steps of the SCoPE2 workflow and suggest alternatives steps and methods for future analyses.

[1] `https://github.com/SlavovLab/SCoPE2`

[2] `https://uclouvain-cbio.github.io/SCP.replication/articles/SCoPE2.html`

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

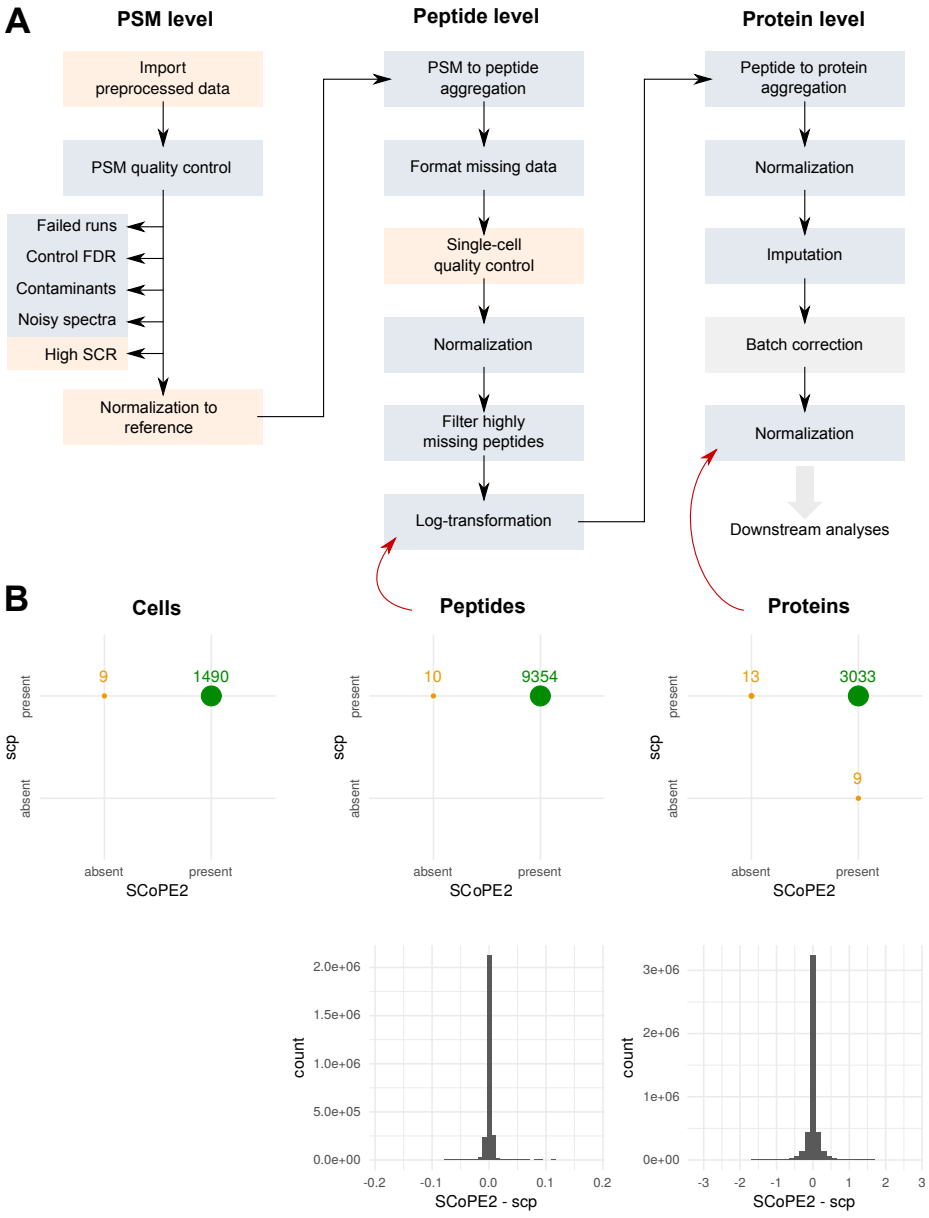

Figure 1: **Reproducing the SCoPE2 data processing. A:** Overview of the key steps performed in the SCoPE2 script. Blue boxes indicate steps that are already implemented in `QFeatures`. The orange boxes indicate steps that were implemented in `scp`. The gray box indicates a step implemented in another package. **B:** Results of the replication. The top row demonstrates the agreement between cells, peptides or proteins between SCoPE2 and `scp`. The bottom row shows the numerical differences between the peptide or protein expression matrices. Red arrows point towards the step that generated the tested data.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4 Conclusion

New tools are required for principled and standardized analysis of SCP data. In this work, we show the successful application of `scp`, our R/Bioconductor software package, to reproduce the data processing workflow published in Specht et al. (2021). While replication or reproduction don't guarantee optimal processing of the data and the results, it demonstrates coherence and increase trust in the data and the results. In addition, the `scp` package allows for an open SCP environment that can foster new methodological developments as well as spreading SCP data analysis towards a broader computational community. We emphasized on the standardization of the implementation which facilitates the integration with currently available tools such as the single-cell methods and workflows provided by the Bioconductor project (Amezquita et al. (2019)). Furthermore, the code is continuously tested and improved to guarantee long term usability of the software.

Although the reproduction of the SCoPE2 results supports the reliability of the original work, additional improvements are necessary. Complex challenges, such as batch effects and data missingness, still need to be tackled and further methodological developments are required for a principled and rigorous workflow.

5 Expert opinion

5.1 Batch correction

The SCoPE2 protocol relies on sample multiplexing. The 1490 single cell samples were multiplexed across 177 MS runs, 63 of which were labeled using TMT-11 and 114 using TMT-16. The data were acquired across 4 chromatographic batches (LCA9, LCA10, LCB3 and LCB7). Unsurprisingly, batch effects account for the main source of variation in the unprocessed peptide data, as indicated by a principal component analysis (PCA) on Figure 2. The first component (12.4 % of total variance) perfectly separates the TMT-11 from the TMT-16 batches and the second component (6 % of total variance) further separates the four chromatographic batches. The next two components (7.3 % of total variance) are driven by biological variations and separate macrophages from monocytes. Because components in PCA are constrained to be orthogonal, this analysis indicates that technical and biological variation are independent. This is a key assumption in order to separate the undesired technical variability from the biological variability. Orthogonality between technical and biological variation is a achieved by a careful design of experiment. As pointed out in the SCoPE2 protocol,

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

it is crucial to randomize cell types and biological samples across different MS batches.

Since batch effects are technically unavoidable, they need to be accounted for computationally. The SCoPE2 authors opted for removing the batch effect using ComBat, an empirical Bayes framework (Johnson et al., 2007). As with any procedure, it is important to understand and apply the requirements of the method. First, ComBat assumes a balanced design, i.e. it requires that differences between batches be only the result of technical differences. This can be an issue when cell types or cell states are unknown in advance, i.e. when the single-cell experiments are designed, such as for the unsupervised discovery of cell populations. Second, ComBat cannot work with missing data which requires the data to be imputed beforehand. As we will discuss later, imputation is a sensitive step that can lead to substantial artifacts in the data, especially when the number of missing values is high, as is the case for single-cell proteomics data. Thirdly, ComBat cannot account for the hierarchical structure of batch effects. We anticipate that once the technology matures, and is applied to clinical samples, for instance across multiple patients and acquisitions, that such a hierarchical structure will become significant. Finally, ComBat creates a new data set by fitting and removing the batch effect from the input data and ignores the uncertainty associated to the estimation of the batch effect itself. It would be important to quantify this uncertainty instead of considering point estimates. Other batch correction methods have been developed for scRNA-Seq data and were extensive benchmarked elsewhere (Tran et al. (2020) and Chazarra-Gil et al. (2021)). However, methods tailored for other single-cell application only partly address the above listed issues and none suggest to propagate the uncertainty linked to batch effect estimation. An alternative approach would be to avoid batch correction altogether and account for batch effects explicitly during data modeling (Ritchie et al., 2015; Goeminne et al., 2016; Risso et al., 2018).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

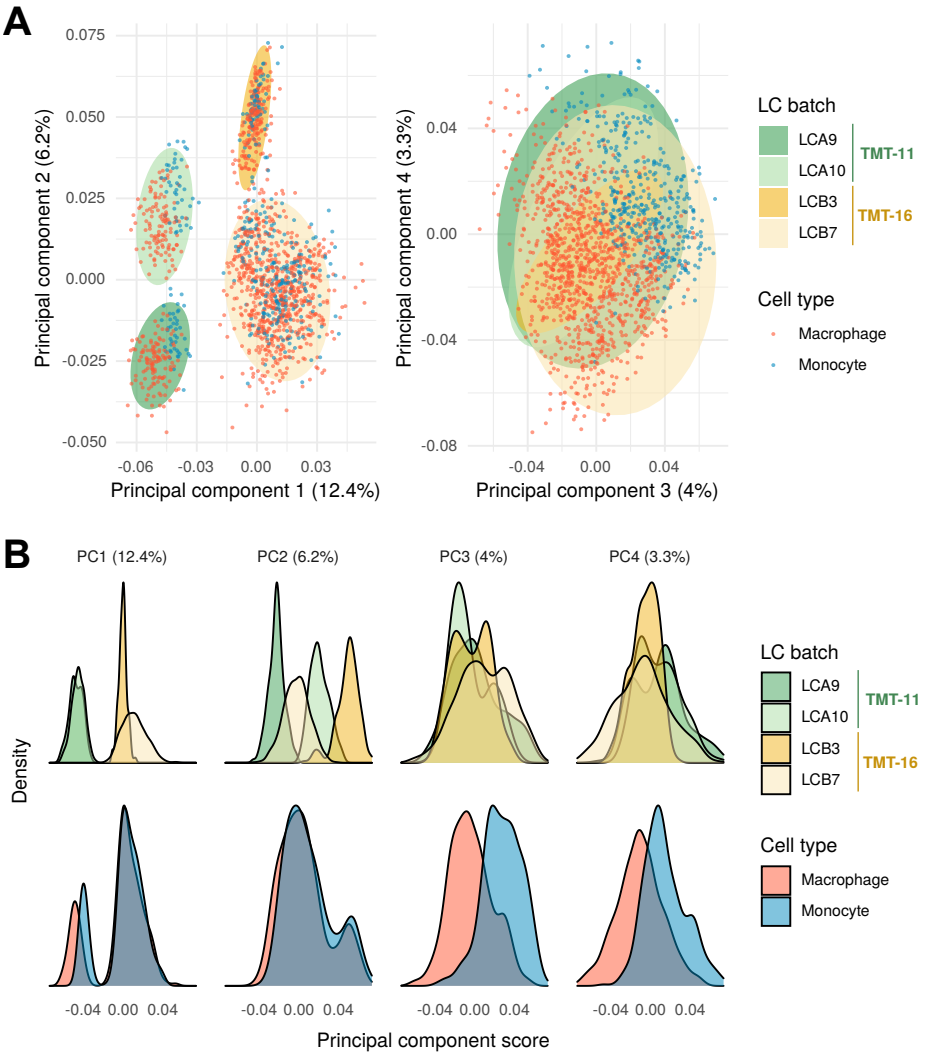

Figure 2: **SCP data exhibit batch effects.** The PCA is performed on the peptide data after log-transformation (*cf* Figure 1A). The Nonlinear Iterative Partial Least Squares (NIPALS) algorithm (Andrecut (2009)) was used to account for missing values during PCA. The LC batches were acquired either with a TMT-11 (green) or TMT-16 (yellow) protocol. The data set contains two types of single-cells: macrophages (red) and monocytes (blue). **A:** PCA scores for the first four components. Each point represents a single-cell is colored according to the corresponding cell type. The ellipses give the 95 % interval for each chromatographic batch. **B:** Distribution of the principal components scores. Each principal component is displayed in a separate column. The distributions are split according to LC batch (top row) or to the sample type (bottom row). The density were computed from the PCA scores.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

5.2 Data missingness

Next to batch effects, missing values are a major challenge in MS-based proteomics Lazar et al. (2016). Missingness refers to the fact that not all features (peptides or proteins) are detected and quantified in all samples. We can distinguish between to types of missingness.

The first type is biological missingness. The peptides of a proteins are not detected in a sample because that sample does not express the protein. Such missingness is biologically relevant and must be considered accordingly. We observed this phenomenon in the SCoPE2 data, where some peptides are systematically missing less in macrophages compared to monocytes and the reduced missingness is correlated with an increased average expression level in that cell type (Figure 3).

The second type is technical missingness. There are several technical mechanisms that explain why a protein could not be detected in a sample. A first reason is that none of its constituting peptides could be correctly delivered to the MS instrument, for example due to sample loss. Sample loss is a major concern for single-cell applications because only limited amounts of material are available to start with. This limitation is actively being tackled and improved in the last two years, the SCoPE2 protocol (Specht et al. (2021)) or the nanoPOTS chips (Zhu et al. (2018b)) are two examples among others. Poor ionization of peptides can also lead to reduced signal or completely missing data. Another cause of technical missingness is related to MS1 peak selection. In data dependent acquisition (DDA), only the most abundant precursor peaks are selected for fragmentation and MS2 acquisition. Whether a peak will be selected is therefore highly dependent on the abundance of a peptide, the surrounding peptides in a given sample, and that peptides ionization efficiency. Several approaches have been developed to reduce this bias by propagating spectrum identifications from one sample to MS1 peaks from another sample. The match between run algorithm of MaxQuant (Tyanova et al. (2016)) is very popular in label-free SCP (Zhu et al. (2018a), Zhu et al. (2018b), Zhu et al. (2019), Cong et al. (2020), Cong et al. (2021), Brunner et al. (2020)), but methodological improvements have recently been suggested for both label-free (Kalxdorf et al. (2020)) and TMT-based SCP (Yu et al. (2020)). Finally, another reason for missingness is the inability to match a spectrum to a peptide sequence. This usually occurs when lowly abundant peptides generate low quality spectra. Therefore, the more abundant a peptide is, the more likely it will get identified. This limitation is tackled by improving the current sensitivity of LC-MS/MS instruments. For instance, Cong et al. (2020) reported an improved proteome coverage when decreasing the diameter of the LC columns or upgrading the Orbitrap Eclipse Tribrid MS to an Orbitrap Fusion Lumos Tribrid MS. Later, they also showed improved peptide identification by coupling the MS with a high field

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

asymmetric ion mobility spectrometry (FAIMS) device (Cong et al. (2021)). Technical missingness translates to the fact that two similar MS runs will not contain the same set of quantified proteins. Although most proteins are common to several MS runs, each run exhibits a specific set of proteins that were probably present but missed in the other runs (Figure 3). While upcoming technical improvements to SCP will further decrease the amount of missing values, computational approaches will still be required.

To overcome the current limitations regarding missing data, Specht and colleagues imputed missing data using the k-nearest neighbors (KNN) method. They applied KNN in the sample space instead of the gene space, thus increasing the similarity between different samples. Since subsequent cluster or differential abundant protein analyses focus on sample-wise differences, this causes an underestimation of the variance and hence leads to a potential increased of false positive outcome.

Furthermore, the imputation is performed at the protein level. As pointed out by Lazar et al. (2016), imputation at protein levels means that a first implicit imputation is performed at the peptide level and the authors suggest to use instead well-justified imputation methods. However, a good understanding of the missingness mechanism is required to justify the use of a suited imputation method. Further research is required to extend the work of Lazar et al. (2016) to the context of SCP data. Finally, just like batch correction, imputation is an estimation process that generates estimates with some degree of uncertainty. Replacing missing data by imputed values ignores the variance associated to the estimates and this variance can become large when available data are scarce. Multiple imputation, i.e. the application of a range of imputation parameters or methods to estimate a range of plausible values rather than point estimates, would be a promising strategy here. This is best illustrated by an issue we noticed in the data. For instance, the RNF41 protein is quantified in only three MS runs and KNN imputation predicted the missing values for the remaining runs (Figure 4). When comparing the resulting data distribution for to the distribution for VIM, a protein that is not missing, we can clearly observe that the imputation introduces two suspicious trends. First, the variability observed for imputed values is much lower than for acquired values, and second, the imputation does not exhibit batch effects. While reduced variability and absence of batch effects are desirable properties, in this case, we are faced with erroneous data that does not hold biologically meaningful information. The imputed data for RNF41 is unreliable and should be flagged accordingly.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

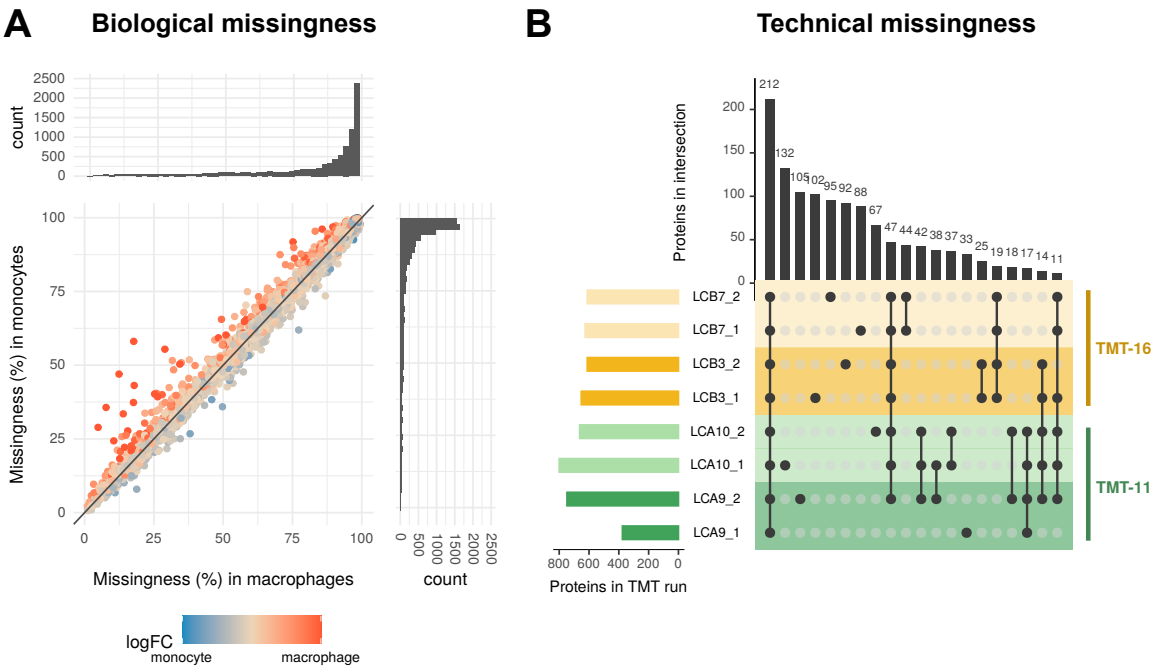

Figure 3: **Missing data is the consequence of two components. A.** Biological missingness is illustrated by plotting the proportion of missing values in monocytes against the proportion of missing values in macrophages for each peptide. Those proportions are also shown on the histograms along the y and x axis for monocyte and macrophage, respectively. Each peptide is colored according the relative log fold change between macrophage over monocyte. The data used is the peptide data after log-transformation (cf Figure 1 A). **B.** The technical missingness is shown using an upset plot (Gehlenborg, 2019) on eight representative MS runs. Two MS runs were randomly sampled from each of the four LC batches. The bar plot on the left shows the total number of proteins per MS run and the bar plot at the top shows the number of proteins for each intersection. A black dot indicates the corresponding MS runs that are included in the intersection.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

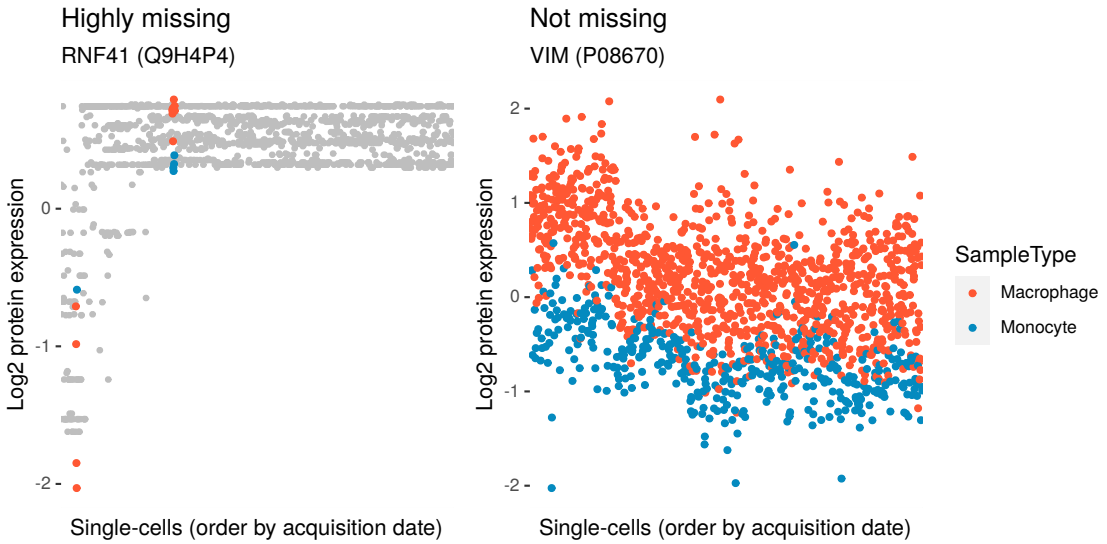

Figure 4: **Problem with imputation.** The data distribution is shown for two proteins: RNF41, a highly missing protein, and VIM, a protein with no missing data. The variance associated to the imputed values for RNF41 (gray points) is not correctly estimated as compared to the variance observed for VIM. Data points are colored in red for macrophages and in blue for monocyte.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

5.3 Batch effects and data missingness are not independent

As of today, all published SCP analyses consider batch effects and data missingness as two distinct issues that can be tackled separately when they are, in reality correlated. Figure 5 highlights the impact of acquiring data in different LC batches on the data missingness. First, since with SCoPE2, peptides are identified from the carrier signal, the number of identified peptides and their missingness display a prominent MS acquisition effect. Second, the LC batches influence the amount of missingness. For instance, more missing values are observed for the batch LCB3 than LCB7. Third, the amount of missing data within each LC batch varies over time. LCA10 displays a very clear increase of missingness, while LCB7 show a decrease is missing values. Finally, LC batches also influences the variability of missing data as the proportion of missing values, with LCB3 displaying much less thereof compared to all other ones. Therefore missing values can only be correctly modeled if we include batch covariates. Inversely, batch effect can only be correctly modeled if we accurately model the missing data instead of replacing by imputed values.

A solution to this issue is to explicitly model the protein expression and the protein detection rate. The hurdle model suggested in Goeminne et al. (2020) is very compelling in this regard. The hurdle model consist of two components. The first component is the MSqRob model (Goeminne et al., 2016), that fits peptide intensities as a function of sample covariates, and includes blocking factors for batch effect, taking into account the correlation between peptides belonging to the same protein. Inference on the estimates allow to perform differential abundance analysis. The second component is a binary component that models the probability that an observation is missing as a function of sample covariates for each run independently. This allows to perform differential detection analysis. Further research is needed to assess the performance of the model when applied to SCP data in the light of inflation of missing values, and to further adapt the algorithm to achieve principled SCP data analysis.

In conclusion, we believe there are two open paths of research that need to be explored to deal with the batch effect and data missingness challenge. First, we need to better understand the different mechanisms that influence missingness and batch effects in SCP data and how they differ from bulk proteomics. Benchmark data sets are therefore required to assess our ability to control for technical factors (e.g. operator, acquisition run, instrument, LC column, ...) while preserving the variance induced by biological meaningful (e.g. cell type, cell state, treatment,...). The second path to take is to develop dedicated analytical models and methods that can disentangle the technical challenges that are batch effects and data missingness from the desired biological

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

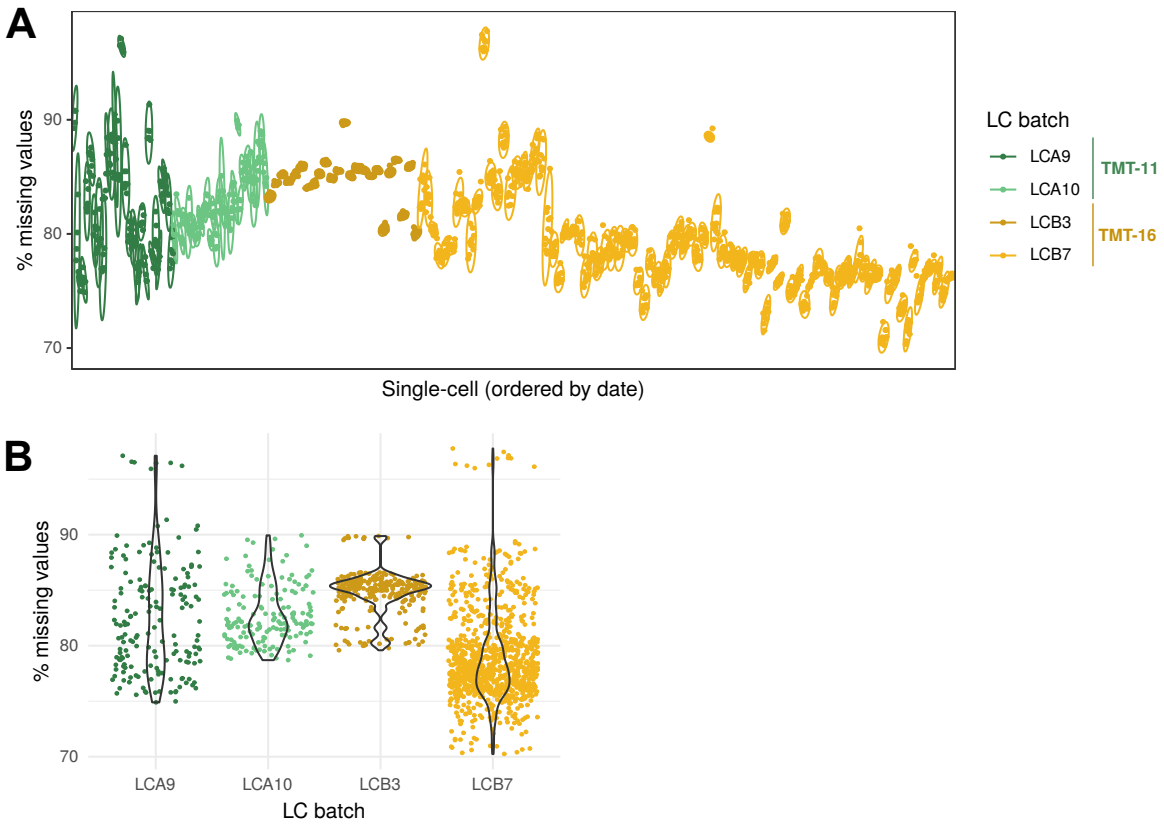

Figure 5: **Influence of batch on data missingness.** The proportion of missing data is shown for each single cell as a dot colored by LC batch. **A.** Effect of the MS run. Cells are ordered based on the acquisition date. The 95 % ellipses are drawn for every MS run. **B.** Effect of LC batch. Cells are grouped by LC batch. The missing data distribution within each batch is highlighted using violin plots.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

knowledge. `scp` represents an ideal environment for a standardized processing of the data and hence allowing comparison, integration and improvement of various existing methods available from other fields as well as benchmarking new methodological innovations.

Acknowledgements Christophe Vanderaa was supported by a PhD fellowship from the Belgian National Fund for Scientific Research (FNRS). The authors would also like to thank Nikolai Slavov, Edward Emmott and Harrison Specht for their openness in sharing all their data and scripts, and responsiveness in addressing questions and comments.

Conflict of interest The authors declare no conflict of interest.

References

Robert A Amezquita, Aaron T L Lun, Etienne Becht, Vince J Carey, Lindsay N Carpp, Ludwig Geistlinger, Federico Martini, Kevin Rue-Albrecht, Davide Risso, Charlotte Soneson, Levi Waldron, Hervé Pagès, Mike L Smith, Wolfgang Huber, Martin Morgan, Raphael Gottardo, and Stephanie C Hicks. Orchestrating single-cell analysis with bioconductor. *Nat. Methods*, pages 1–9, December 2019.

M Andrecut. Parallel GPU implementation of iterative PCA algorithms. *J. Comput. Biol.*, 16(11): 1593–1599, November 2009.

Andreas-David Brunner, Marvin Thielert, Catherine Vasilopoulou, Constantin Ammar, Fabian Coscia, Andreas Mund, Ole B Horning, Nicolai Bache, Amalia Apalategui, Markus Lubeck, Oliver Raether, Melvin A Park, Sabrina Richter, David S Fischer, Fabian J Theis, Florian Meier, and Matthias Mann. Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. December 2020.

Ruben Chazarra-Gil, Stijn van Dongen, Vladimir Yu Kiselev, and Martin Hemberg. Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. *Nucleic Acids Res.*, February 2021.

Yongzheng Cong, Yiran Liang, Khatereh Motamedchaboki, Romain Huguet, Thy Truong, Rui Zhao, Yufeng Shen, Daniel Lopez-Ferrer, Ying Zhu, and Ryan T Kelly. Improved Single-Cell proteome

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

coverage using Narrow-Bore packed NanoLC columns and ultrasensitive mass spectrometry. *Anal. Chem.*, January 2020.

Yongzheng Cong, Khatereh Motamedchaboki, Santosh A Misal, Yiran Liang, Amanda J Guise, Thy Truong, Romain Huguet, Edward D Plowey, Ying Zhu, Daniel Lopez-Ferrer, and Ryan T Kelly. Ultrasensitive single-cell proteomics workflow identifies >1000 protein groups per mammalian cell. *Chem. Sci.*, 12(3):1001–1006, 2021.

Claudia Ctortecka and Karl Mechtler. The rise of single-cell proteomics. *Analytical Science Advances*, n/a(n/a), February 2021.

Maowei Dou, Geremy Clair, Chia-Feng Tsai, Kerui Xu, William B Chrisler, Ryan L Sontag, Rui Zhao, Ronald J Moore, Tao Liu, Ljiljana Pasa-Tolic, Richard D Smith, Tujin Shi, Joshua N Adkins, Wei-Jun Qian, Ryan T Kelly, Charles Ansong, and Ying Zhu. High-Throughput single cell proteomics enabled by multiplex isobaric labeling in a nanodroplet sample preparation platform. *Anal. Chem.*, 91(20):13119–13127, October 2019.

Laurent Gatto. QFeatures: Quantitative features for mass spectrometry data, 2020.

Nils Gehlenborg. *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*, 2019. URL `https://CRAN.R-project.org/package=UpSetR`. R package version 1.4.0.

Ludger J E Goeminne, Kris Gevaert, and Lieven Clement. Peptide-level robust ridge regression improves estimation, sensitivity, and specificity in data-dependent quantitative label-free shotgun proteomics. *Mol. Cell. Proteomics*, 15(2):657–668, February 2016.

Ludger J E Goeminne, Adriaan Sticker, Lennart Martens, Kris Gevaert, and Lieven Clement. MSqRob takes the missing hurdle: Uniting intensity- and Count-Based proteomics. *Anal. Chem.*, 92(9):6278–6287, May 2020.

Wolfgang Huber, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper D Hansen, Rafael A Irizarry, Michael Lawrence, Michael I Love, James MacDonald, Valerie Obenchain, Andrzej K Oleś, Hervé Pagès, Alejandro Reyes, Paul Shannon, Gordon K Smyth, Dan Tenenbaum, Levi Waldron, and Martin Morgan. Orchestrating high-throughput genomic analysis with bioconductor. *Nat. Methods*, 12(2):115–121, February 2015.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, January 2007.

Mathias Kalxdorf, Torsten Müller, Oliver Stegle, and Jeroen Krijgsveld. IceR improves proteome coverage and data completeness in global and single-cell proteomics. November 2020.

Ryan T Kelly. Single-Cell proteomics: Progress and prospects. *Mol. Cell. Proteomics*, August 2020.

Cosmin Lazar, Laurent Gatto, Myriam Ferro, Christophe Bruley, and Thomas Burger. Accounting for the multiple natures of missing values in Label-Free quantitative proteomics data sets to compare imputation strategies. *J. Proteome Res.*, 15(4):1116–1125, April 2016.

Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.*, 9(1):284, January 2018.

Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, 43(7):e47, April 2015.

Erwin M Schoof, Nicolas Rapin, Simonas Savickas, Coline Gentil, Eric Lechman, James Seymour Haile, Ulrich auf Dem Keller, John E Dick, and Bo T Porse. A quantitative Single-Cell proteomics approach to characterize an acute myeloid leukemia hierarchy. August 2019.

Nikolai Slavov. Unpicking the proteome in single cells. *Science*, 367(6477):512–513, January 2020.

Nikolai Slavov. Single-cell protein analysis by mass spectrometry. *Curr. Opin. Chem. Biol.*, 60:1–9, February 2021.

Harrison Specht and Nikolai Slavov. Transformative opportunities for Single-Cell proteomics. *J. Proteome Res.*, 17(8):2565–2571, August 2018.

Harrison Specht, Edward Emmott, Aleksandra A Petelski, R Gray Huffman, David H Perlman, Marco Serra, Peter Kharchenko, Antonius Koller, and Nikolai Slavov. Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biol.*, 22(1):50, January 2021.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.12.439408; this version posted April 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.*, 21(1):12, January 2020.

Stefka Tyanova, Tikira Temu, and Juergen Cox. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.*, 11(12):2301–2319, December 2016.

Christophe Vanderaa and Laurent Gatto. *Single cell replication package*, 2021. URL `https://uclouvain-cbio.github.io/SCP.replication/index.html`.

Sung-Huan Yu, Pelagia Kyriakidou, and Jürgen Cox. Isobaric matching between runs and novel PSM-Level normalization in MaxQuant strongly improve reporter Ion-Based quantification. *J. Proteome Res.*, 19(10):3945–3954, October 2020.

Ying Zhu, Maowei Dou, Paul D Piehowski, Yiran Liang, Fangjun Wang, Rosalie K Chu, William B Chrisler, Jordan N Smith, Kaitlynn C Schwarz, Yufeng Shen, Anil K Shukla, Ronald J Moore, Richard D Smith, Wei-Jun Qian, and Ryan T Kelly. Spatially resolved proteome mapping of laser capture microdissected tissue with automated sample transfer to nanodroplets. *Mol. Cell. Proteomics*, 17(9):1864–1874, September 2018a.

Ying Zhu, Paul D Piehowski, Rui Zhao, Jing Chen, Yufeng Shen, Ronald J Moore, Anil K Shukla, Vladislav A Petyuk, Martha Campbell-Thompson, Clayton E Mathews, Richard D Smith, Wei-Jun Qian, and Ryan T Kelly. Nanodroplet processing platform for deep and quantitative proteome profiling of 10-100 mammalian cells. *Nat. Commun.*, 9(1):882, February 2018b.

Ying Zhu, Mirko Scheibinger, Daniel Christian Ellwanger, Jocelyn F Krey, Dongseok Choi, Ryan T Kelly, Stefan Heller, and Peter G Barr-Gillespie. Single-cell proteomics reveals changes in expression during hair-cell development. *Elife*, 8, November 2019.