BIOINFORMATICS

Comparative analysis of methods for batch correction in proteomics — a two-batch case

Katerina Danko[1], Lavrentii Danilov[2], Anna Malashicheva[3], and Arseniy Lobov[3]

[1]Bioinformatics Institute, ul. Kantemirovskaya, 2, Saint Petersburg, 197342, Russian Federation
[2]Department of Genetics and Biotechnology, Faculty of Biology, Saint Petersburg State University, Universitetskaya nab., 7–9, Saint Petersburg, 199034, Russian Federation
[3]Laboratory of Regenerative Biomedicine, Institute of Cytology, Russian Academy of Sciences, Tikhoretskiy pr., 4, Saint Petersburg, 194064, Russian Federation

Address correspondence and requests for materials to Arseniy Lobov, lobov@incras.ru

Abstract

A proper study design is vital for life science. Any effects unrelated to the studied ones (batch effects) should be avoided. Still, it is not always possible to exclude all batch effects in a complicated omics study. Here we discuss an appropriate way for analysis of proteomics data with an enormous technical batch effect. We re-analyzed the published dataset (PXD032212) with two batches of samples analyzed in two different years. Each batch includes control and differentiated cells. Control and differentiated cells form separate clusters with 209 differentially expressed proteins (DEPs). Nevertheless, the differences between the batches were higher than between the cell types. Therefore, the analysis of only one of the batches gives 276 or 290 DEPs. Then we compared the efficiency of five methods for batch correction. ComBat was the most effective method for batch effect correction, and the analysis of the corrected dataset revealed 406 DEPs.

Keywords: batch effect, proteomics, bioinformatics, batch effect correction

Introduction

In an experimental or comparative study, two or more sets of objects that differ in some specific factors are compared, for example, a pea with yellow or green seeds. Researchers should avoid any factors, unrelated to the studied one, but able to cause systematic differences between the samples. Such effects are called "batch effects". They may have biological nature, e. g., donor effect. However, they are usually associated with technical issues such as differences in instruments or reagents used for various samples (Goh et al., 2017). The problem is that such batch effects in the data may dramatically decrease the sensitivity of the statistical analysis. Thus, we should give them a lot of attention.

In technically complicated omics studies, systemic differences in dozens of technical factors may cause batch effects. There are several reviews addressing appropriate study design to avoid batch effects, but it is not always possible to exclude all of them (Goh and Wang, 2017). For example, in standard Illumina RNA-seq transcriptomics, all samples should be analyzed in the same cell while in mass-spectrometry-based proteomics all samples should be analyzed randomly in the same run. Beyond that, all samples should be prepared using the same reagents, plastic, and instruments. Variation in operators is also important. As a result, it is impossible to exclude all batch effects and sometimes the researcher does not even know about batch effects in the data. From that point, it is crucial to be able to detect possible batch effects and deal with them.

Specific methods for batch effect recognition and correction were created. Initially, these methods were developed for microarray and, then, for next-generation

Citation: Danko, K., Danilov, L., Malashicheva, A., and Lobov, A. 2023. Comparative analysis of methods for batch correction in proteomics — a two-batch case. Bio. Comm. 68(1): 56–61. https://doi.org/10.21638/spbu03.2023.106

Authors' information: Katerina Danko, PhD Student, orcid.org/0000-0003-3987-2175; Lavrentii Danilov, PhD Student, orcid.org/0000-0002-4479-3095; Anna Malashicheva, Dr. of Sci. in Biology, Head of Laboratory, orcid.org/0000-0002-0820-2913; Arseniy Lobov, PhD, Senior Researcher, orcid.org/0000-0002-0930-1171

Manuscript Editor: Pavel Skutschas, Department of Vertebrate Zoology, Faculty of Biology, Saint Petersburg State University, Saint Petersburg, Russia

Received: September 19, 2022;

Revised: November 27, 2022;

Accepted: January 13, 2023.

Funding: The work was supported by the Russian Science Foundation for Basic Research, project No. 18-14-00152.

Ethics statement: This paper does not contain any studies involving human participants or animals performed by any of the authors.

Competing interests: The authors have declared that no competing interests exist.

sequencing data. For the last decade batch effect correction methods have been actively adapted for proteomics studies (Čuklina et al., 2021). One common strategy for batch effect correction is to create linear gene expression models within biological groups (e. g., disease and control groups, cell types) and confounding effects (e. g., batch labeling, patient identifiers) as covariates. One of the first implementations of the ComBat method (Johnson, Li, and Rabinovic, 2007) used an empirical Bayesian algorithm. The modern version of ComBat-seq uses a negative binomial regression approach (Zhang, Parmigiani, and Johnson, 2020). There are also modern approaches based on different methods of clustering, quantile normalization, and machine learning (Kiselev et al., 2017; Fei and Yu, 2020; Shaham et al., 2017).

Mass-spectrometry-based proteomics has several specific features which might be addressed during batch effect correction. First is the problem of inferring the ratio of peptides to proteins (Rosenberger et al., 2014; Teo et al., 2015; Muntel et al., 2019). Because of the specificity of protein intensity (their number is determined by the number of measured peptides or even fragment ions), a decision has to be made at which level the data will be adjusted. Another problem in proteomic studies is that MS signal drift can occur (Jiang et al., 2020). To compensate for the influence of these factors, there is a large number of statistical methods for data correction. Ideally, the data should be adjusted at the MS-fragment level (Čuklina et al., 2021) by specific approaches. To date, however, most methods have been developed for RNA-seq data rather than proteomics data. Only recently, due to the increase in proteomics data, researchers have become concerned about creating such approaches (Čuklina et al., 2021).

In this article we aimed to compare the different techniques for batch effect correction and recognition developed for both proteomics and transcriptomics data, using experimental data devoted to the study of the mechanisms of osteogenic differentiation. We used a dataset, deposited in the ProteomExchange consortium with the dataset identifier PXD032212. The dataset was originally obtained during proteomics comparative analysis of molecular mechanisms of osteogenic differentiation of human valve interstitial cells (VICs) isolated from healthy donors or patients with calcific aortic valve disease (Semenova et al., 2022). Due to the high value and rarity of human material, they performed two shotgun proteomics analyses of VICs from two donor groups in 2018 and 2019. All cells were prepared in the same labs and two shotgun proteomics launches were performed with the same instrument in the same conditions. Nevertheless, the instrument state cannot be the same after a year of active exploitation and, as was mentioned by the authors of the original study (Semenova et al., 2022), there is strong batch effect. Thus, this study indicates a good example of the study design with an enormous technical batch effect which has no obvious connection with the investigated biological differences. From our experience, such a study design problem is relatively common, but there are not so many examples of an extended discussion of such datasets and appropriate ways for their analysis.

Therefore, here we compared five popular methods for batch effect correction and demonstrated that adequate correction of a batch effect significantly increases the sensitivity of further statistical analysis and the number of identified DEPs compared to the analysis of uncorrected data or separate analysis of each batch.

Material and methods

Data availability statement

The article contains no new data. Fully reproducible code for the data analysis is deposited on the Github repository (https://github.com/kvdanko/Batch-effect-correction-methods).

Protein identification

Firstly, we reanalyzed the presented shotgun proteomics data. Protein and peptide identification was performed in Peaks Xpro software (Bioinformatics Solutions Inc.) using human protein sequences from the SwissProt database (uploaded on 2.03.2021; 20,394 protein sequences). The following search parameters were applied: parent mass error tolerance of 15 ppm, fragment mass error tolerance of 0.05 ppm, protein and peptide FDR 1 %, trypsin protease, and two possible missed cleavage sites. Proteins with at least two unique peptides were included into further analysis. Cysteine carbamidomethylation was set as a fixed modification. Methionine oxidation, acetylation of protein N-term, asparagine, and glutamine deamidation were set as variable modifications.

Data analysis

The PEAKS output was used for further statistical analysis in R (version 3.6.3). In the first steps, we removed proteins with missed values in more than two samples and performed the imputation of missed values by k-nearest neighbors by the "impute" package and quantile normalization by the "limma" package. Then we used five different methods to eliminate the batch effect (Table 1) which are based on different statistical approaches: sva (Leek et al., 2021), bapred (Hornung and Causeur, 2016), limma (Ritchie et al., 2015), Harman (Oytam et al., 2016).

To choose the optimal batch correction method, we used principal component analysis (PCA), guided PCA (gPCA), partial least squares-discriminant analysis

Table 1. List of statistical methods with their descriptions and R package version which were tested to eliminate the batch effect

Method	Description	R package
ComBat	Empirical Bayes method	sva (version 3.32.1)
Harman	Based on PCA. Reduces batch effect and keeps user-defined class effects	Harman (version 1.12.0)
Ratio A	Ratio-based method scaling the expression values by the arithmetic mean	bapred (version 1.0)
Ratio G	Ratio-based method scaling the expression values by the geometric mean	bapred (version 1.0)
BMC (batch mean centering)	Centering the variables within batches to have zero mean	limma (version 3.40.6)

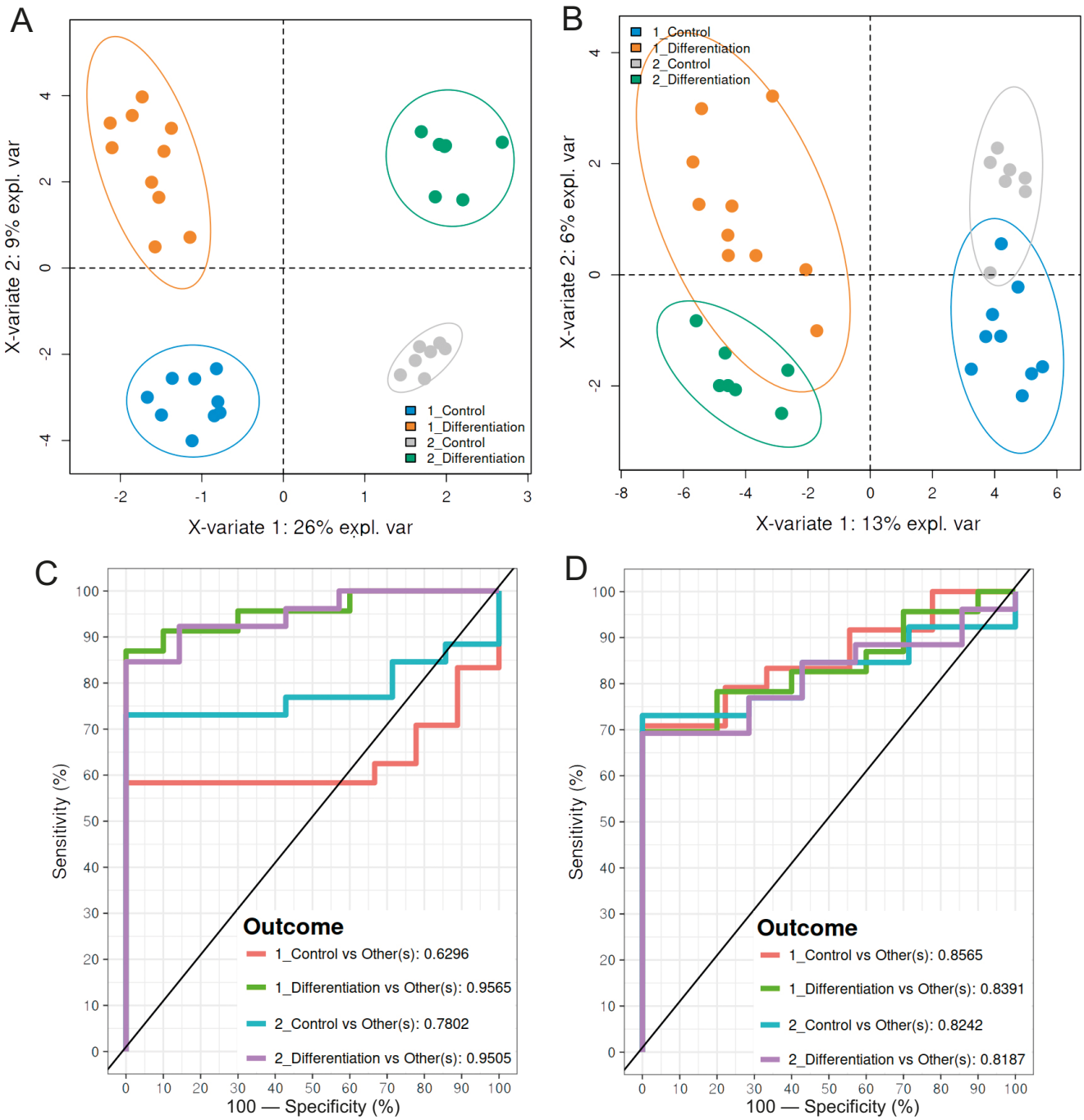

Fig. 1. Comparison of partial least squares-discriminant analysis (PLS-DA) of the dataset of proteomics analysis of human valve interstitial cells (VICs) before and after batch effect correction by ComBat (SVA). 1_Control — control VICs analyzed in 2018; 1_Differentiation — differentiated VICs analyzed in 2018; 2_Control — control VICs analyzed in 2019, 2_Differentiation — differentiated VICs analyzed in 2019. (A, B) PLS-DA score plot of the dataset of control and differentiated VICs analyzed in 2018 or 2019 without (A) or with (B) batch effect correction. (C–D) ROC curves corresponding to PLS-DA analysis of dataset with (C) or without (D) batch effect correction.

(PLS-DA; package MixOmics; Rohart, Gautier, Singh, and Lê Cao, 2017), and differential expression analysis (package limma). The main task was to reveal the most effective methods which would remove batch effects but allow to save as much biological variation as possible.

Results and discussion

After data filtration, 1207 proteins were included into further analysis. We found a strong batch effect in the data which far exceeds biological differences between control and differentiated VICs. It is clearly visible on PLS-DA (Fig. 1A) where four clear clusters could be seen (control and differentiated VICs from the first or second year) with two variants associated with dispersion between years (x-variate 1: 26 % explained variance) and with differentiation (x-variate 2: 9 % explained variance). A similar pattern was observed on PCA (Fig. 2A). Normally, we should see separate clusters for different experimental conditions (healthy and disease, control and differentiated, etc.) and no other distinguished clusters on PCA or sample correlation plot. If we see clear clusters unrelated to the experimental conditions like in Fig. 2, it should probably be regarded as a batch effect.

There are also specific batch effect detection methods such as gPCA. According to Fig. 2, we also found a sig-

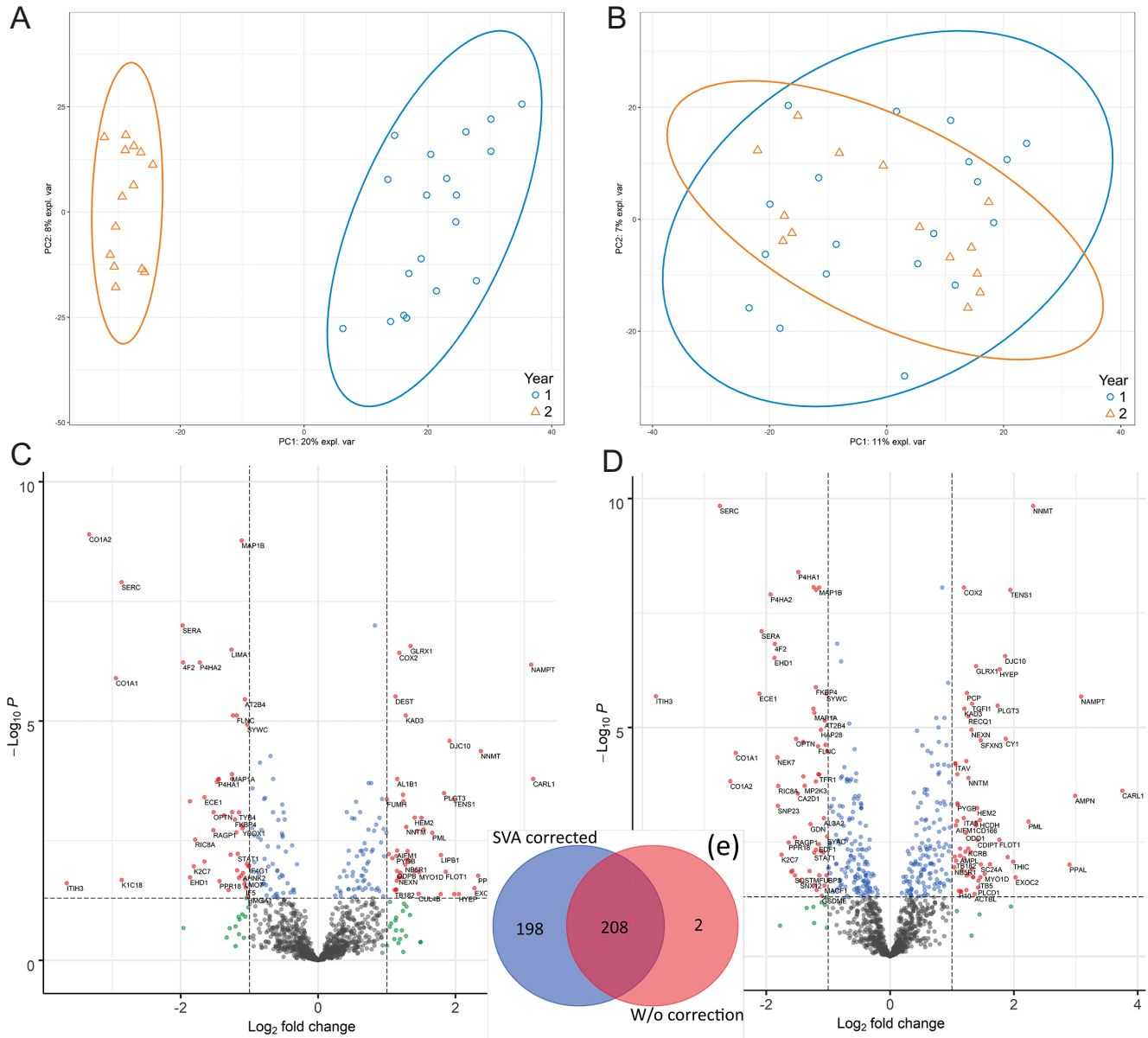

Fig. 2. Comparison of clusterization and sensitivity of differential expression analysis in the dataset of proteomics analysis of human valve interstitial cells (VICs) with or without batch effect correction by ComBat (SVA). (A, B) PCA plot with ellipses marked clusters corresponding to two proteomics analysis performed in two different years of two VICs cohorts (two batches) before (A) and after (B) batch correction. (C–E) Comparison of results of differential expression analysis between control and differentiated VICs. (C) Vulcanoplot of differentially expressed proteins without correction. (D) Vulcanoplot of differentially expressed proteins after correction of batch effect by ComBat (SVA). (E) Venn diagram representing comparison of statistically significant differentially expressed proteins found in dataset with or without batch effect correction by ComBat (SVA).

BIO-INFORMATICS

nificant batch effect by gPCA (Table 2). Obviously, such strong technical variation could mask the studied biological differences between undifferentiated and differentiation cells. Nevertheless, 210 differentially expressed proteins were identified between control and differentiated VICs by limma (Fig. 2C). We assume that appropriate removal of the emphasized batch effect might enhance the sensitivity of differential expression analysis.

Table 2. Results of guided principal components analysis (gPCA) for original dataset without batch correction and with correction by five methods described in the main text

Correction method	gPCA delta	p-value
No correction	0.992	<0.001
ComBat	0.074	1
BMC	0.004	1
Ratio A	0.009	1
Ratio G	0.383	0.801
Harman	0.316	0.905

The comparison of all five methods revealed that only three of them were able to completely remove batch effects according to gPCA: ComBat, BMC (limma), and Ratio A (Table 2). All three methods also keep biological variation between control and differentiation VICs visible on PCA or sPLS-DA (data not shown). Among them, ComBat was considered as the most effective method as it demonstrated the differences between control and differentiated VICs. Still, it is only slightly better than BMC (limma) and Ratio A (bapred).

Despite the absence of the batch effect on gPCA, we can still identify some small differences between the two batches in PLS-DA (Fig. 1B) where they form partially overlapping clusters; while we see completely union clusters on PCA (Fig. 2B).

To test whether the removal of batch effects by ComBat is relevant to enhance the sensitivity of further statistical analysis we compared the results of differential expression analysis between control and differentiated cells performed by limma in the original dataset without batch effect correction and after correction by ComBat (Fig. 2C–E). After ComBat correction, we found 406 differentially expressed proteins (Fig. 2C) against only 210 identified in the dataset without correction (Fig. 2D). But the most important point is that 208 out of 210 differentially expressed proteins found in the original dataset were also found after ComBat correction — batch effect correction by ComBat allowed us to identify twice more DEPs without loss of DEPs which might be identified without correction.

The popular choice in study design with enormous batch effect is to include only one of the batches in the analysis. Therefore, we also compared our results of the differential expression analysis with the same analysis performed in the two batches separately — in both batches, we identified fewer DEPs compared to data after batch effect correction, but more than in the mixed dataset without correction: 276 DEPs for the batch of 2018 and 290 DEGs for 2019.

Conclusion

We can conclude that the modern method for batch effect correction significantly enhances the sensitivity of data analysis in datasets with even enormous technical batch effects. Most importantly, if the batch effect is not associated with the biological hypothesis (each batch has all biological groups compared), combining two technical batches gives higher statistical power than analyzing only one batch. As batch effects might be unpredictable even for the researcher, we recommend looking carefully at possible batch effects in the data and removing them in the way we demonstrated here. In our specific case, ComBat was the most effective method for batch effect correction, but we recommend comparing several methods in each specific case.

Acknowledgements

The authors are grateful to professor Jarle Vaage's group which participated in obtaining the original data. The authors express their gratitude to the Institute of Bioinformatics for the productive discussion and helpful advice. Proteomics data analysis was performed in the resource center "Development of Molecular and Cell Technologies", at St. Petersburg State University.

References

Čuklina, J., Lee, C. H., Williams, E. G., Sajic, T., Collins, B. C., Rodríguez Martínez, M., Sharma, V. S., Wendt, F., Goetze, S., Keele, G. R., and Wollscheid, B. 2021. Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial. *Molecular Systems Biology* 17(8):10240. https://doi.org/10.15252/msb.202110240

Fei, T. and Yu, T. 2020. scBatch: batch-effect correction of RNA-seq data through sample distance matrix adjustment. *Bioinformatics* 36(10):3115–3123. https://doi.org/10.1093/bioinformatics/btaa097

Goh, W. W. B., Wang, W., and Wong, L. 2017. Why batch effects matter in omics data, and how to avoid them. *Trends in Biotechnology* 35(6):498–507. https://doi.org/10.1016/j.tibtech.2017.02.012

Hornung, R. and Causeur, D. 2016. bapred: Batch effect removal and addon normalization (in phenotype prediction using gene data). *Stanford. Department of Statistics: Technical Reports*. No. 19.

Jiang, F., Liu, Q., Li, Q., Zhang, S., Qu, X., Zhu, J., Zhong, G., and Huang, M. 2020. Signal drift in liquid chromatography tandem mass spectrometry and its internal standard calibration strategy for quantitative analysis. *Analytical Chemistry* 92(11):7690–7698. https://doi.org/10.1021/acs.analchem.0c00633

Johnson, W. E., Li, C., and Rabinovic, A. 2007. Adjusting batch effects in microarray expression data using empirical

Bayes methods. *Biostatistics* 8(1):118–127. https://doi.org/10.1093/biostatistics/kxj037

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., and Hemberg, M. 2017. SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods* 14(5):483–486. https://doi.org/10.1038/nmeth.4236

Leek, J. T., Johnson, W. E., Parker, H. S., Fertig, E. J., Jaffe, A. E., Zhang, Y., Storey, J. D., and Torres, L. C. 2021. sva: Surrogate variable analysis. 2020. R package version, 3(0).

Muntel, J., Kirkpatrick, J., Bruderer, R., Huang, T., Vitek, O., Ori, A., and Reiter, L. 2019. Comparison of protein quantification in a complex background by DIA and TMT workflows with fixed instrument time. *Journal of Proteome Research* 18(3):1340–1351. https://doi.org/10.1021/acs.jproteome.8b00898

Oytam, Y., Sobhanmanesh, F., Duesing, K., Bowden, J. C., Osmond-McLeod, M., and Ross, J. 2016. Risk-conscious correction of batch effects: maximising information extraction from high-throughput genomic datasets. *BMC Bioinformatics* 17(1):1–17. https://doi.org/10.1186/s12859-016-1212-5

Ritchie, M. E., Phipson, B., Wu, D. I., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. 2015. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43(7):e47–e47. https://doi.org/10.1093/nar/gkv007

Rohart, F., Gautier, B., Singh, A., and Lê Cao, K. A. 2017. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology* 13(11):e1005752. https://doi.org/10.1371/journal.pcbi.1005752

Rosenberger, G., Ludwig, C., Röst, H. L., Aebersold, R., and Malmström, L. 2014. aLFQ: an R-package for estimating absolute protein quantities from label-free LC-MS/MS proteomics data. *Bioinformatics* 30(17):2511–2513. https://doi.org/10.1093/bioinformatics/btu200

Semenova, D., Zabirnyk, A., Lobov, A., Boyarskaya, N., Kachanova, O., Uspensky, V., Zainullina, B., Denisov, E., Gerashchenko, T., Kvitting, J. P. E., and Kaljusto, M. L. 2022. Multi-omics of in vitro aortic valve calcification. *Frontiers in Cardiovascular Medicine* 9. https://doi.org/10.3389/fcvm.2022.1043165

Shaham, U., Stanton, K. P., Zhao, J., Li, H., Raddassi, K., Montgomery, R., and Kluger, Y. 2017. Removal of batch effects using distribution-matching residual networks. *Bioinformatics* 33(16):2539–2546. https://doi.org/10.1093/bioinformatics/btx196

Teo, G., Kim, S., Tsou, C. C., Collins, B., Gingras, A. C., Nesvizhskii, A. I., and Choi, H. 2015. mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. *Journal of Proteomics* 129:108–120. https://doi.org/10.1016/j.jprot.2015.09.013

Zhang, Y., Parmigiani, G., and Johnson, W. E. 2020. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics* 2(3):lqaa078. https://doi.org/10.1093/nargab/lqaa078