A principled approach to process, analyse and interpret single-cell proteomics data

Laurent Gatto

de Duve Institute, UCLouvain

3 October 2024

Computational Systems Biology of Cancer, Paris, October 2024

Mass spectrometry-based single-cell proteomics (SCP) has become a credible player in the single-cell biology arena. Continuous technical improvements have pushed the boundaries of sensitivity and throughput. However, the computational efforts to support the analysis of these complex data have been missing. Strong batch effects coupled to high proportions of missing values complicate the analysis, causing strong entanglement between biological and technical variability. We propose a simple, yet powerful approach to address this need: linear models. We use linear regression to model and remove undesired technical factors while retaining the biological variability, even in the presence of high proportions of missing values. The key advantage of linear models lies in the interpretability of the results they generate. Inspired by previous research, we streamlined modelling and exploration of the patterns induced by known technical and biological factors. The exploration enables a thorough assessment of the model coefficients, and highlights key factors influencing SCP experiments. Further exploration of the unmodelled variance recovers unknown but biologically relevant patterns in the data, leveraging the power of single-cell proteomics technologies. We successfully applied our approach to a diverse collection of SCP datasets, and could demonstrate that it is also amenable for integrating datasets acquired using different technologies. Our approach represents a turning point for principled SCP data analysis, moving the tension point from how to perform the analysis to result generation and interpretation.

Note that the course lecture needs to be didactic and suitable for Master, PhD and Post-Doc level.

Outline

Single-cell technologies unravel cellular heterogeneity

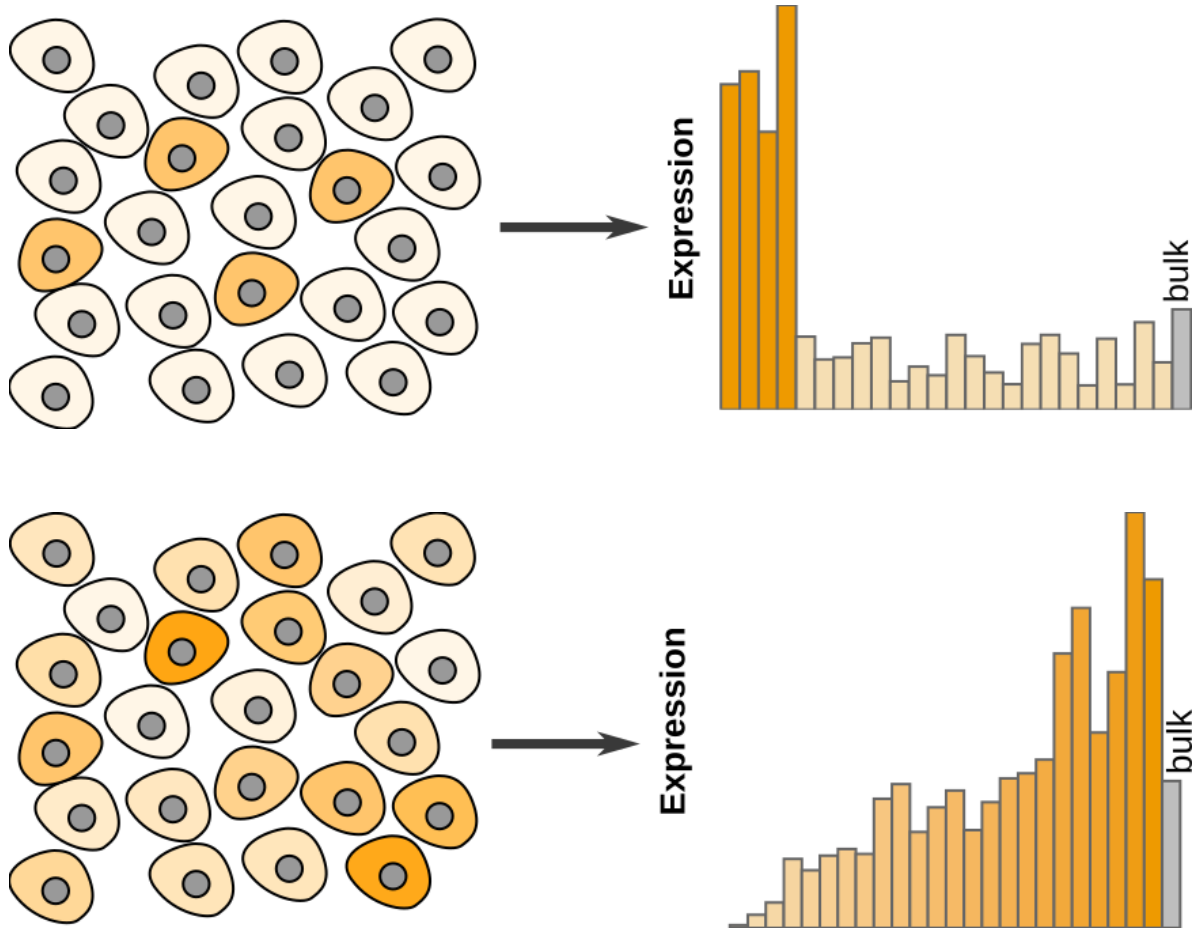
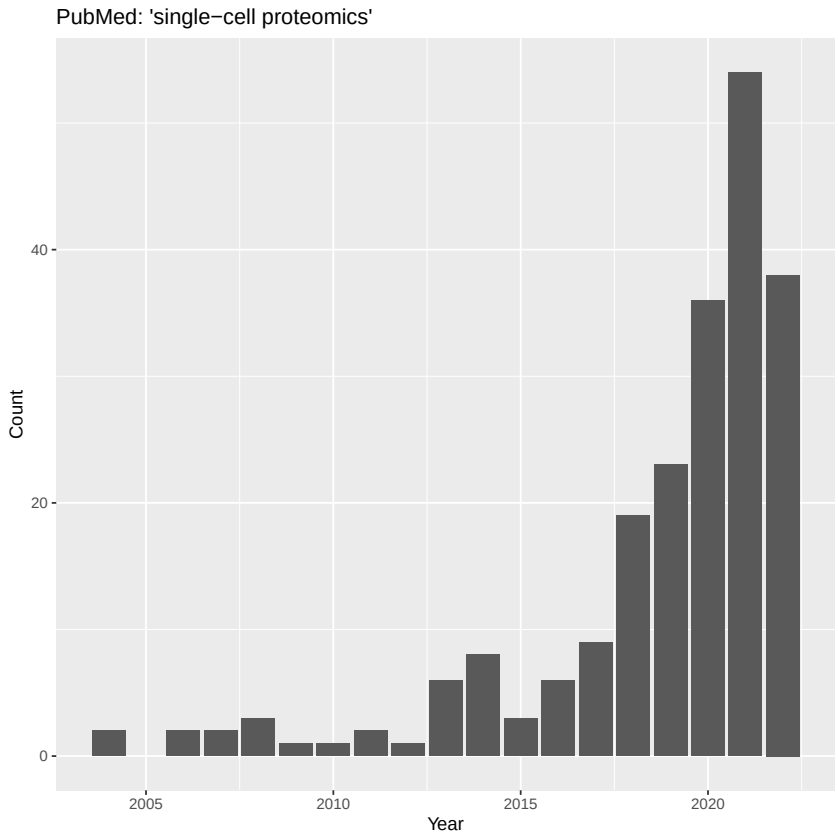


Figure: Cell types and cell states, subpopulation identification, differentiation trajectories (in the absence of known markers).

SCP: . . . – 2019 – 2023 – . . .



August 2019: in a Nature Methods Technology Feature[a], Vivien Marx *dreamt* of single-cell proteomics.

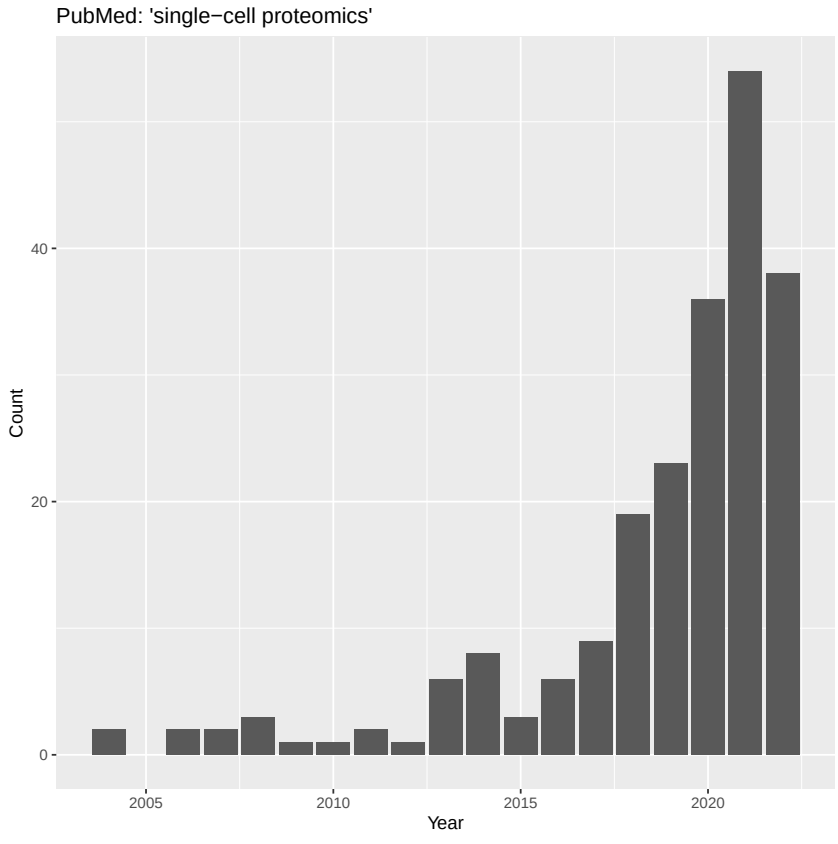
March 2023: Nature Methods published a special issue with a Focus on single-cell proteomics[b] and Initial recommendations for performing, benchmarking and reporting single-cell proteomics experiments[c].

---

[a] 10.1038/s41592-019-0540-6

[b] www.nature.com/collections/bdfhafhdeb

[c] 10.1038/s41592-023-01785-3



Possible through better sample preparation, reduction of loss of material, miniaturisation, automation, better MS, greater sensitivity, DDA and DIA, LFQ and labelling, . . .

. . . and appropriate **experimental designs** and **computational approaches**.

Single-cell technologies

	FC	scRNA-Seq	SCP
features	10	10^4	10^3
cells	10^6	10^4	10^3
samples	10 - 100	1 - 10	1 ...
	sample/cell throughput	feature throughput	functional

Single-cell proteomics

	FC	scRNA-Seq	SCP
features	10	10^4	10^3
cells	10^6	10^4	10^3
samples	10 - 100	1 - 10	1 . . .
	sample/cell throughput	feature throughput	**functional**

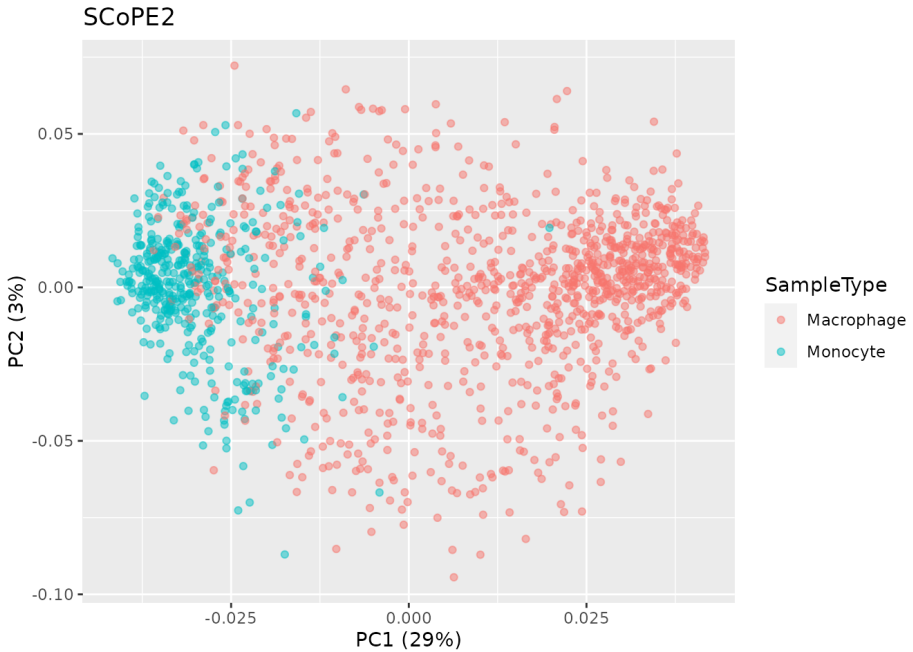
- RNA $\rightarrow$ intention vs. Protein $\rightarrow$ action
- Inference of direct regulatory interactions with minimal assumptions (Slavov, 2022; Hu et al., 2023).
- Post-translational modifications

Outline

Material (1)

The SCoPE2 dataset

- Seminal dataset published by Specht et al. (2021)
- 1096 macrophages, 394 monocytes (after QC)
- 9354 peptides, 3042 proteins
- **Pre-print, data and code available since 2019**



Reproducible research

First steps

- SCoPE2 (and other) repetition/reproduce/replication → **QFeatures** and **scp** packages
- SCoPE2 (and other) data curation → **scpdata** package

More details

- `https://bioconductor.org/packages/QFeatures`
- `https://bioconductor.org/packages/scp`
- `https://bioconductor.org/packages/scpdata`

Build expertise and improve current state-of-the-art

Methods

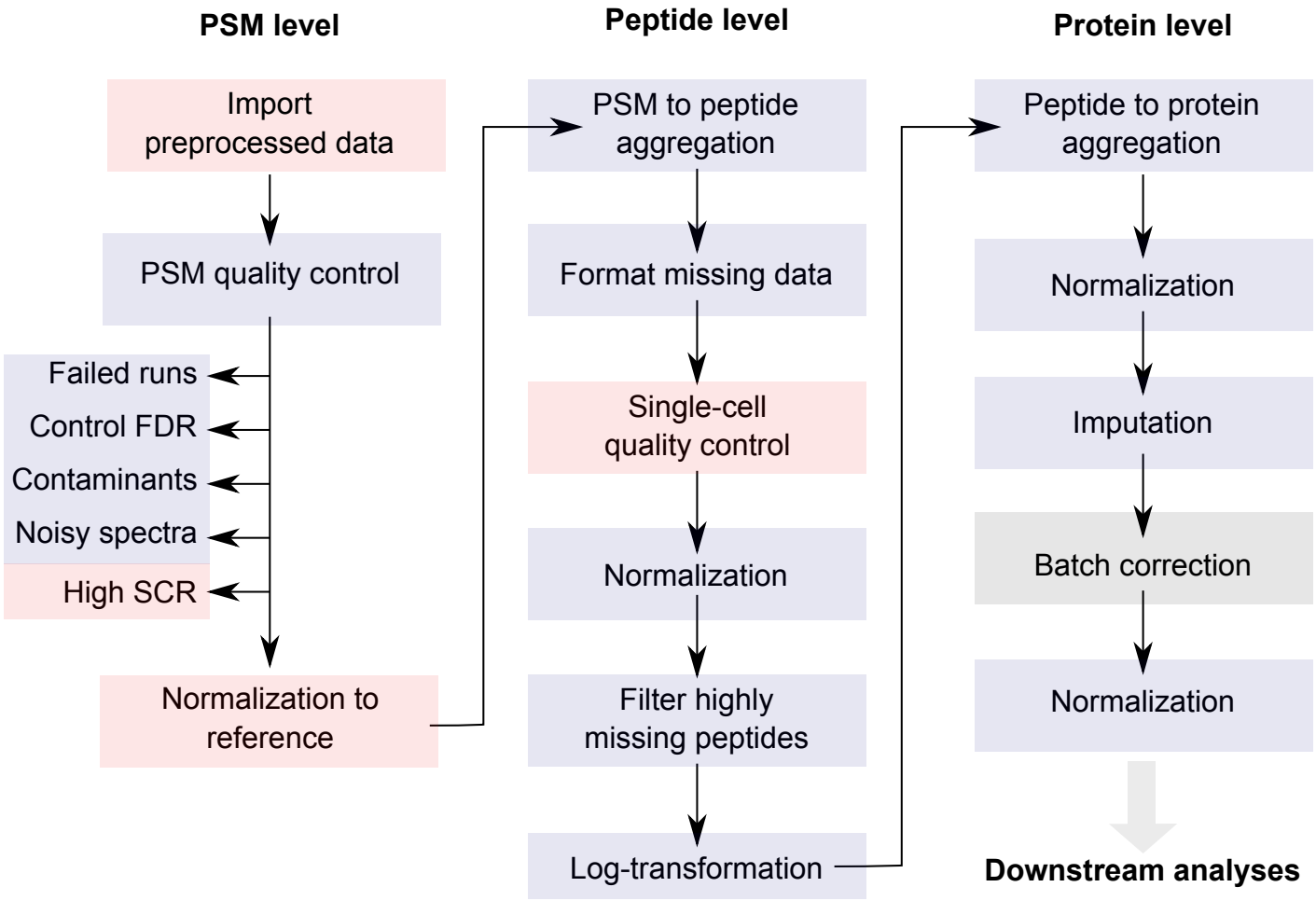


Figure: **Overview of the key steps performed in the SCoPE2 pipeline** (Vanderaa and Gatto, 2021). Blue boxes: **QFeatures**. Red boxes: **scp**. Gray box: **sva::ComBat**.

Outline

Challenge 1: batch effects

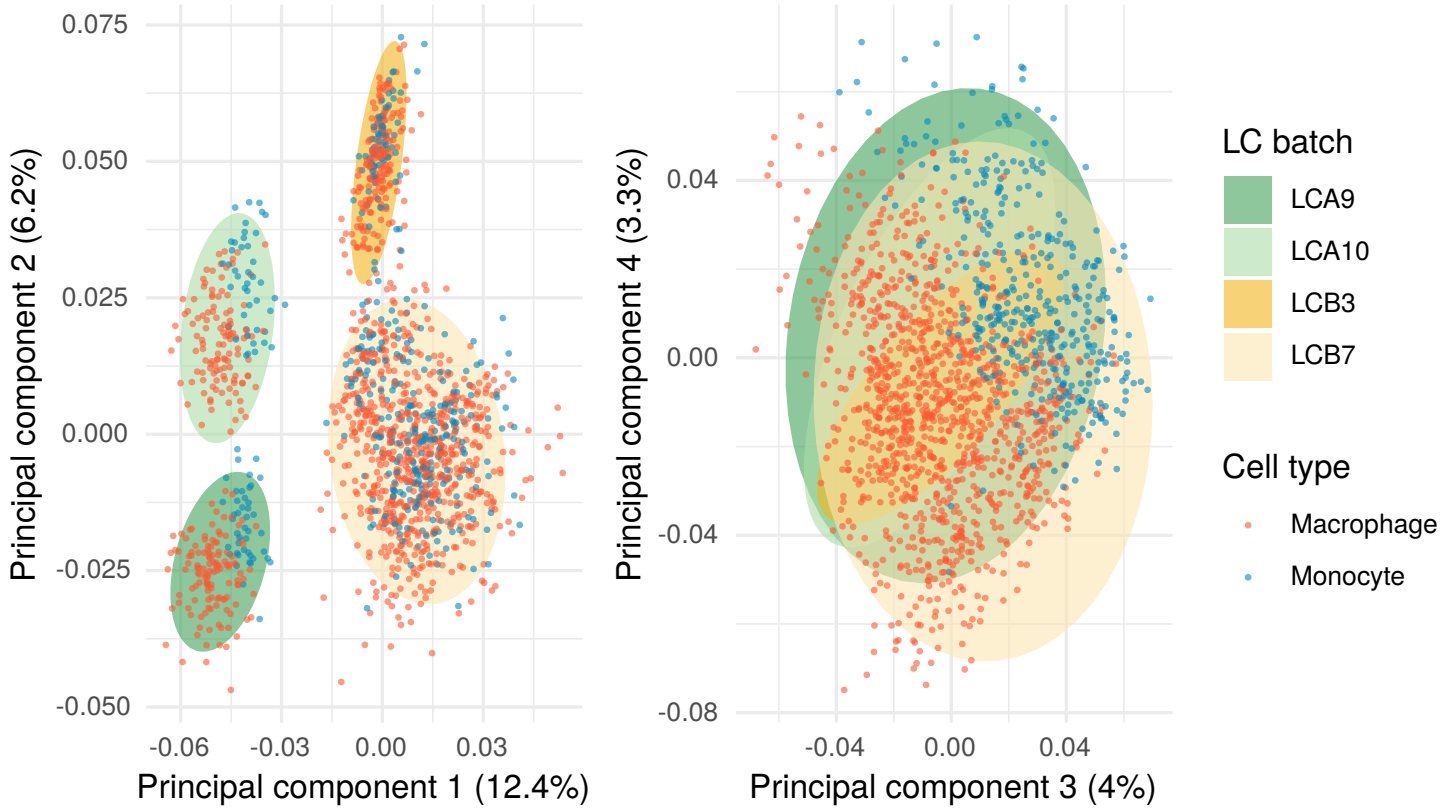


Figure: PCA for the first four components. Each point represents a single-cell and is colored according to the corresponding cell type (Vanderaa and Gatto, 2021).

Challenge 2: missing data

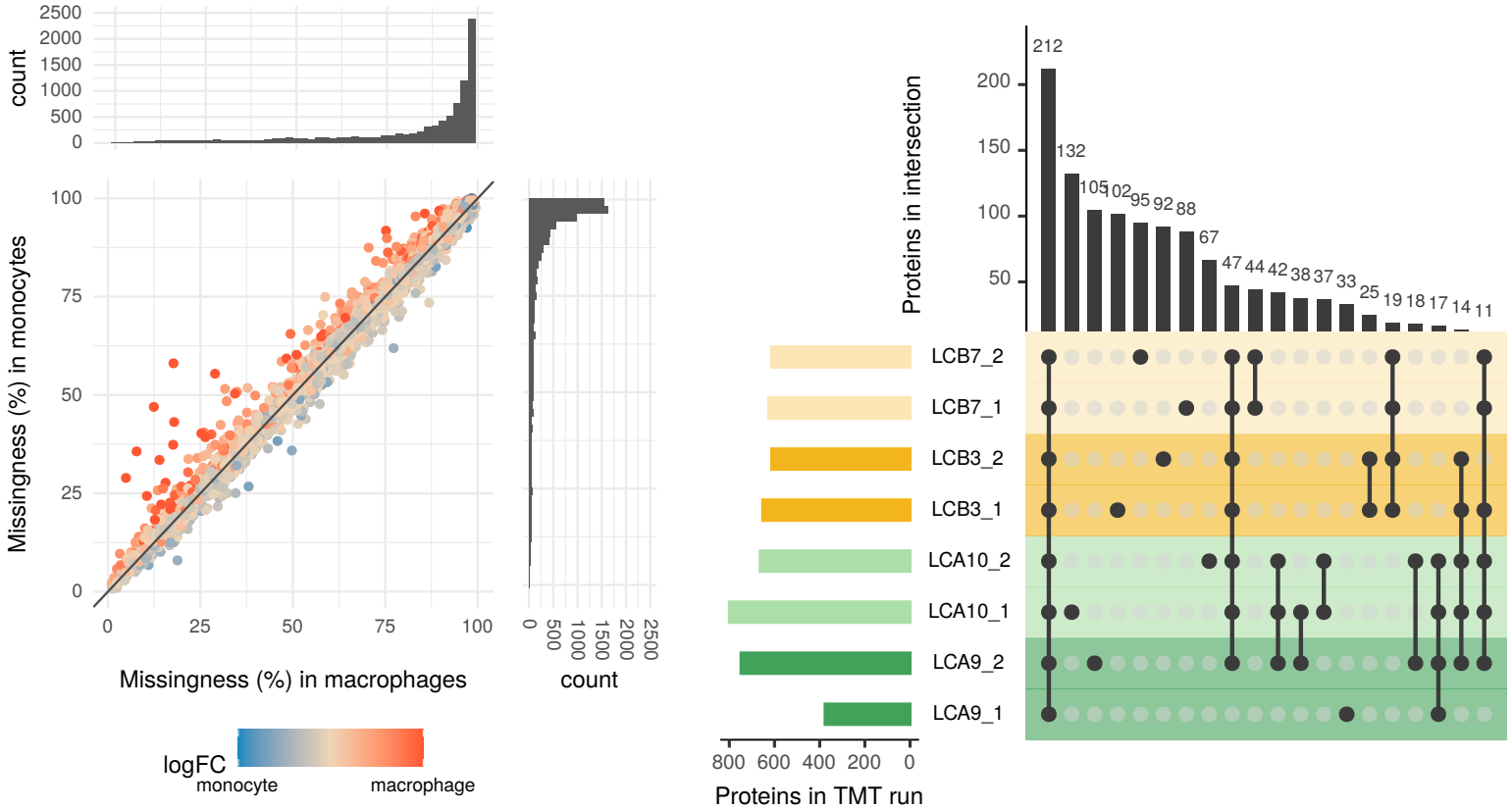


Figure: Missing data is the consequence of biological and technical components (Vanderaa and Gatto, 2021, 2023b).

Challenge 3: 1 + 2

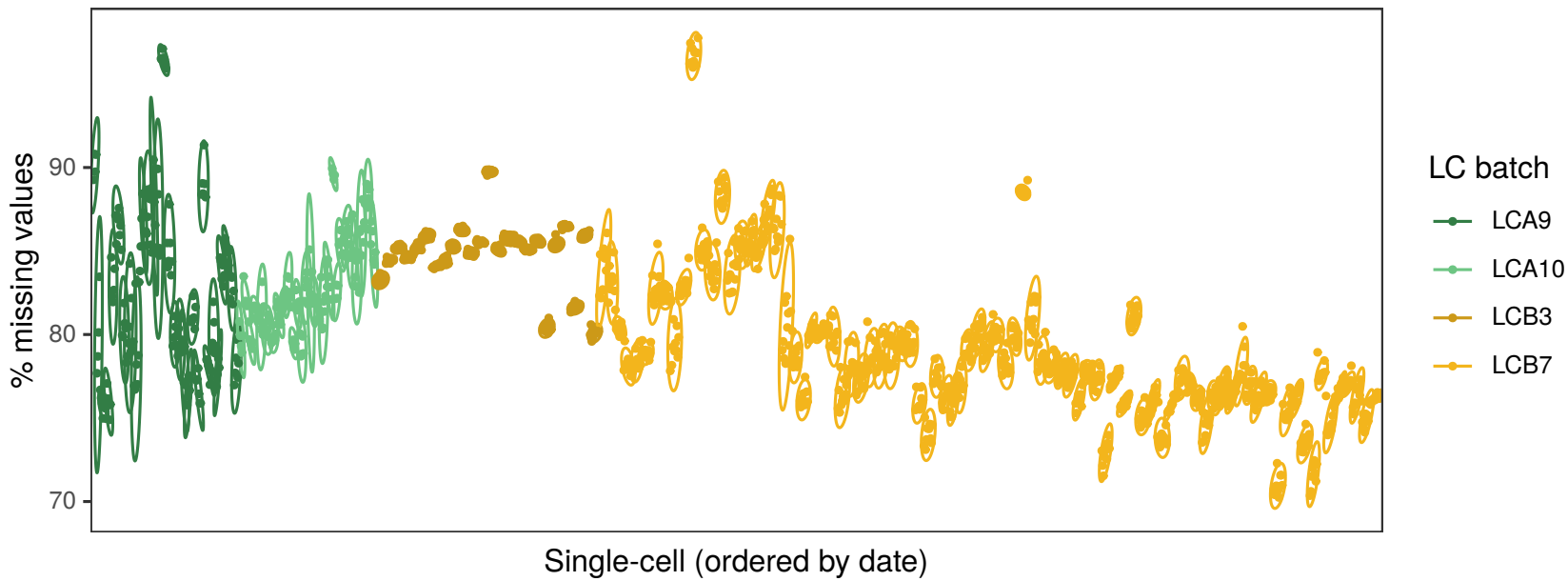
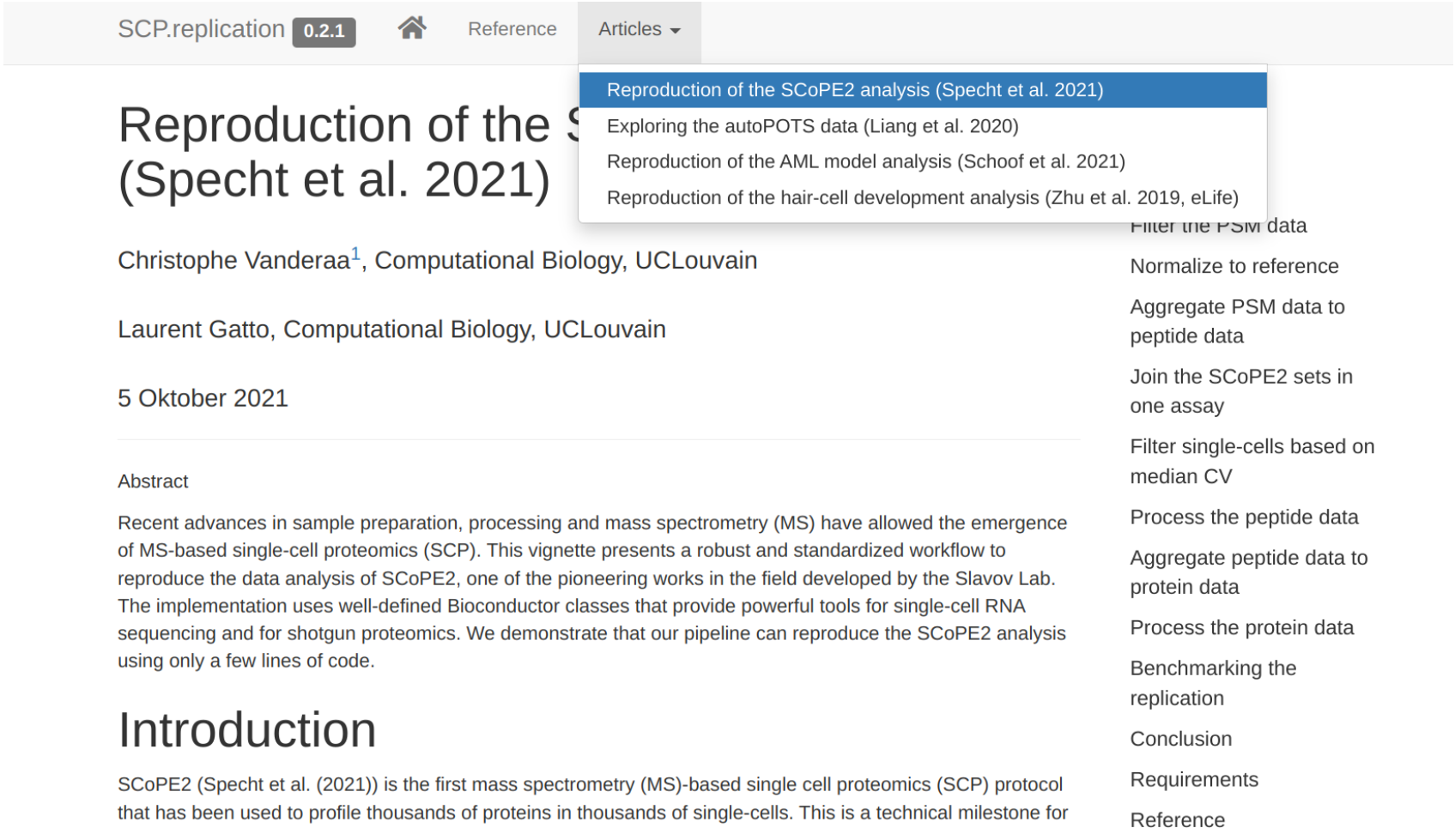


Figure: Influence of batch on data missingness (Vanderaa and Gatto, 2021).

Data analyses review

- How do researches process their data?
- How do they deal with batch effects?
- How do they deal with missing data?

Replication



https://uclouvain-cbio.github.io/SCP.replication

Figure: **SCP.replication**: **systematic reproduction/replication** of published SCP studies using the **scp** package (Vanderaa and Gatto, 2023a).

Systematic review

Figure: Single-cell data processing: **one workflow per paper/lab**. (Vanderaa and Gatto, 2023a)

Problem

- Complex data, many alternative pipelines.
- **Different pipelines produce different results** (see Vanderaa and Gatto (2023a)).
- Little control/understanding of the implications of what is done to the data.

Problem

- Complex data, many alternative pipelines.
- **Different pipelines produce different results** (see Vanderaa and Gatto (2023a)).
- Little control/understanding of the implications of what is done to the data.

Solution: a principled approach

- KISS (*Keep it simple stupid!*), as simple as possible.
- Use what we know to **model** our data.
- Control what we do, **quantify** effects.

Outline

Given that we aren't sure about the effect of data processing...

Given that we aren't sure about the effect of data processing...

Let's start with **minimally processed data**

- Remove low quality precursors and cells
- Aggregate from precursors into peptides
- $\log_2$-transform
- Remove features with *too many* NAs
- No imputation

Given that we aren't sure about the effect of data processing...

Let's start with **minimally processed data**

- Remove low quality precursors and cells
- Aggregate from precursors into peptides
- $\log_2$-transform
- Remove features with *too many* NAs
- No imputation

And use ANOVA–simultaneous component analysis (ASCA)-like methods (Thiel et al., 2017).

(1) Linear modelling

$$y = \beta_0 + \beta_1 \times group + \epsilon$$

$$y = \beta_0 + \beta_1 \times group + \beta_i \times batch_i + \epsilon$$

(2) Quantify the effects' contributions

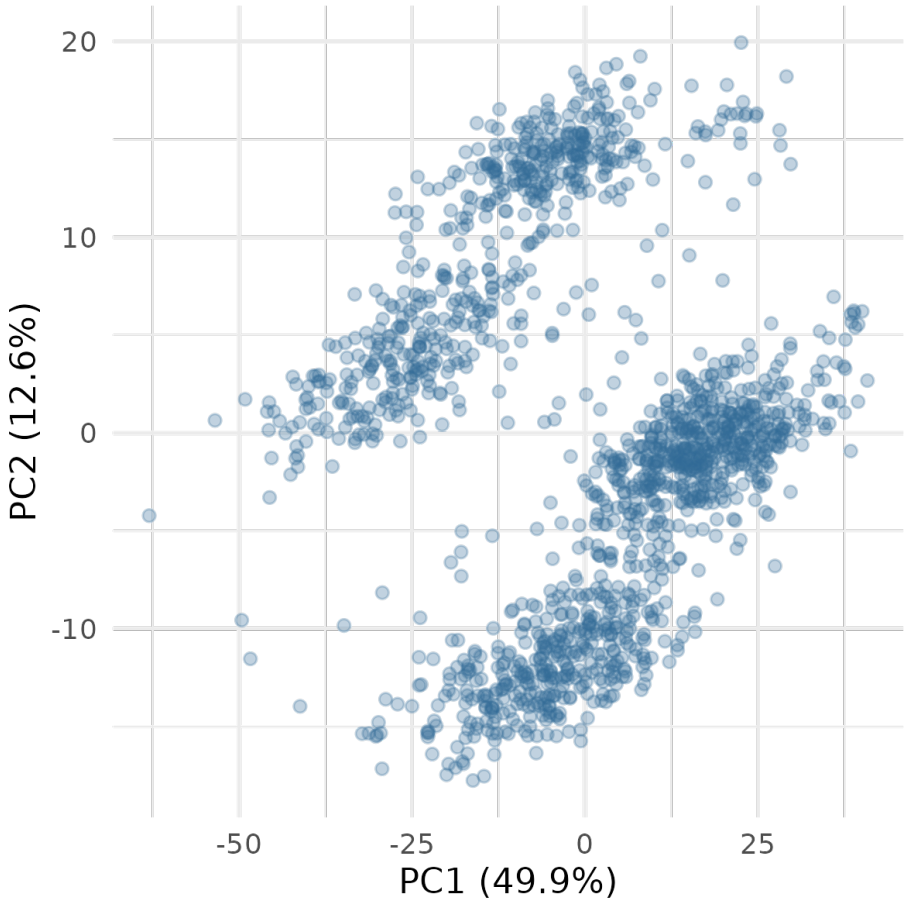
(3) Principal Component Analysis

On effect + residual matrices (of dimensions *features* $\times$ *samples*).

Material (2)

The nPOP dataset

- Data from Leduc et al. (2022)
- nano-ProteOmic sample Preparation
- 877 monocytes, 878 melanoma cells
- 19374 peptides, 3348 proteins
- **Availability of data and code**



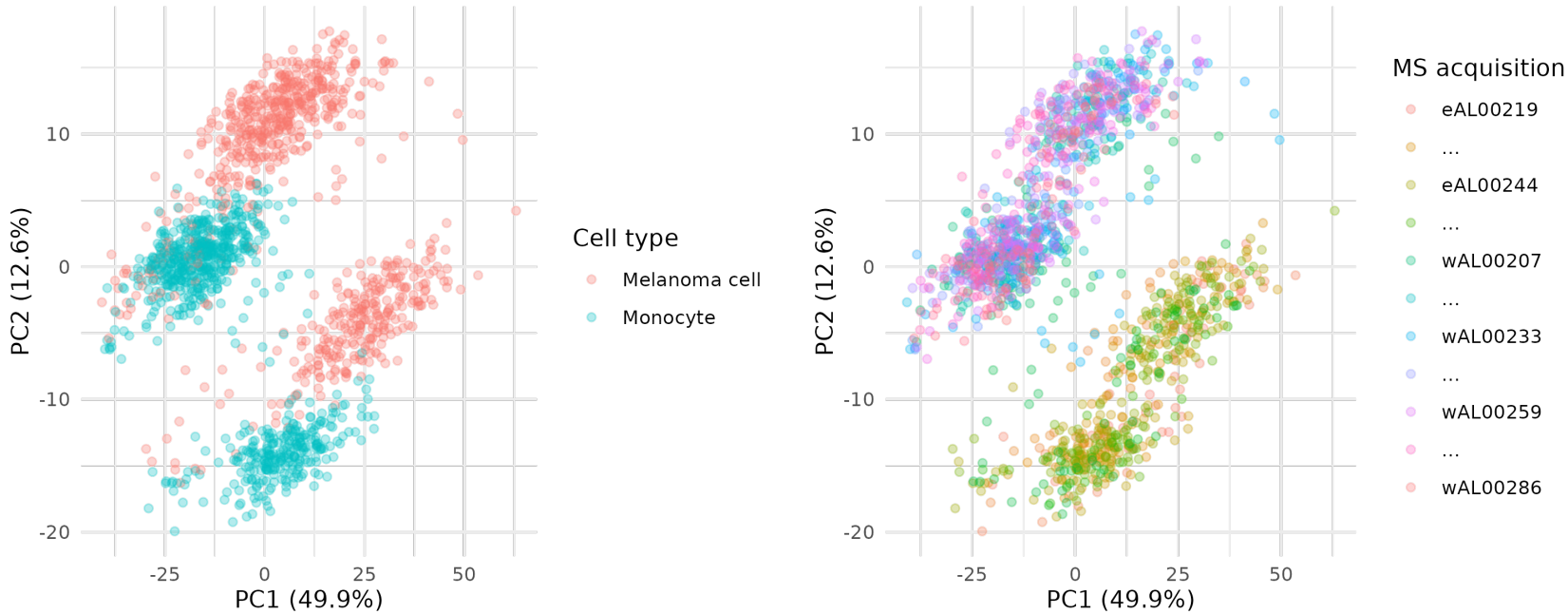


Figure: Melanoma cells and monocytes (left) acquired across multiple acquisition batches (right) (Leduc et al., 2022).

$$y = MS\ \ acquisition + TMT\ \ channel + Cell\ \ type + \epsilon$$

$$y = MS\ acquisition + TMT\ channel + Cell\ type + \epsilon$$

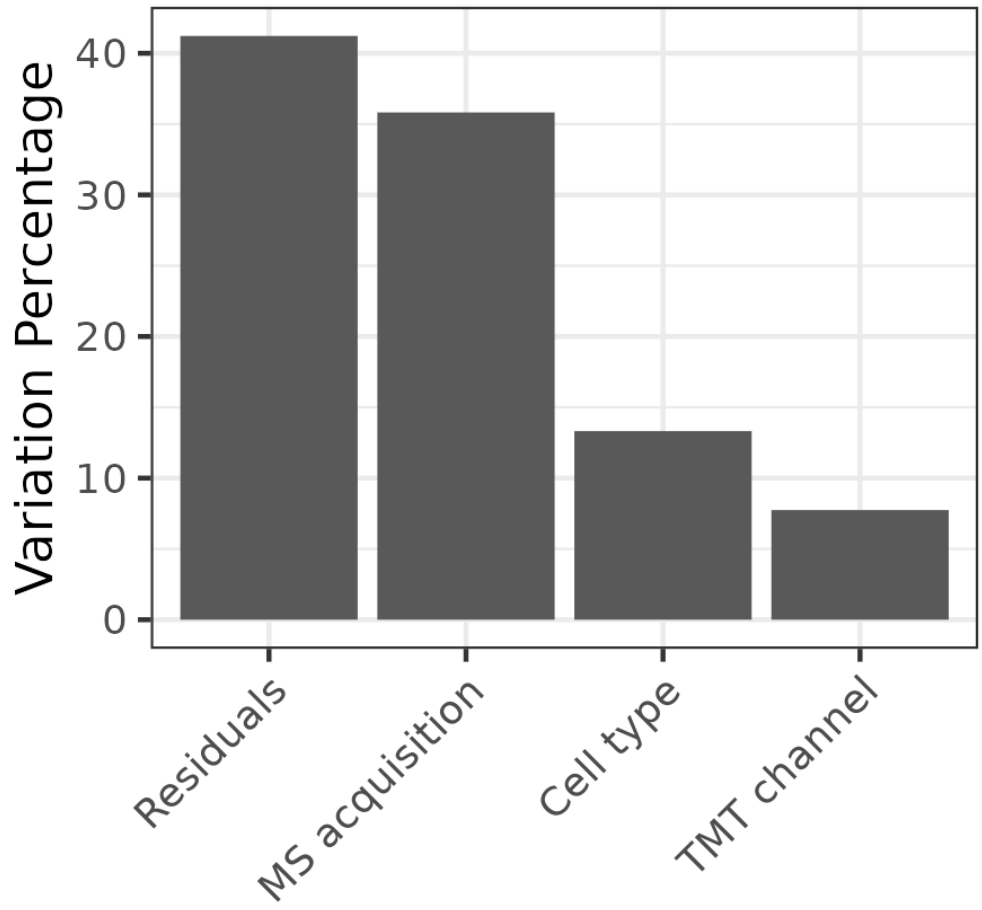
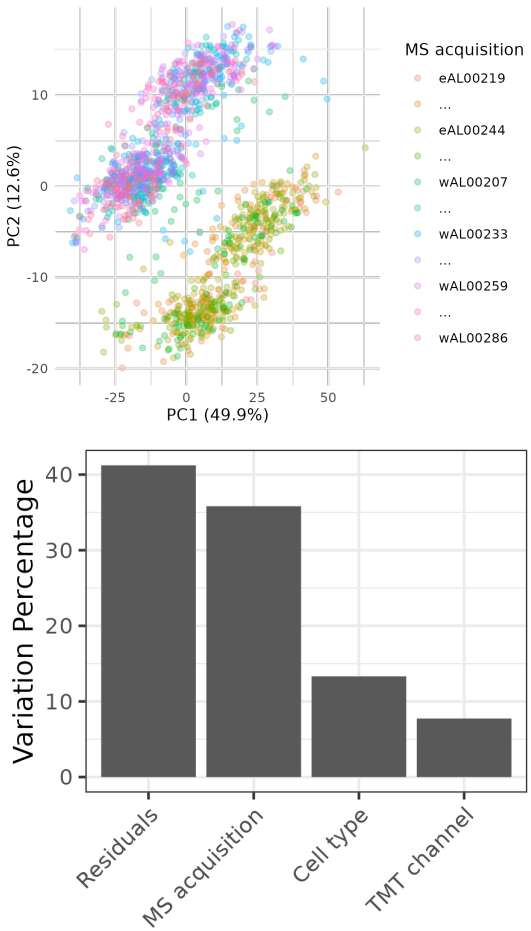


Figure: We are now in a position to **quantify known and unknown effects**: percentages of explained variances of our explained (known) and unexplained (residuals) effects. NB: low biological variance $\neq$ low quality!

PCA on effect matrices

$$y = \textit{MS acquisition} + \textit{TMT channel} + \textit{Cell type} + \epsilon$$



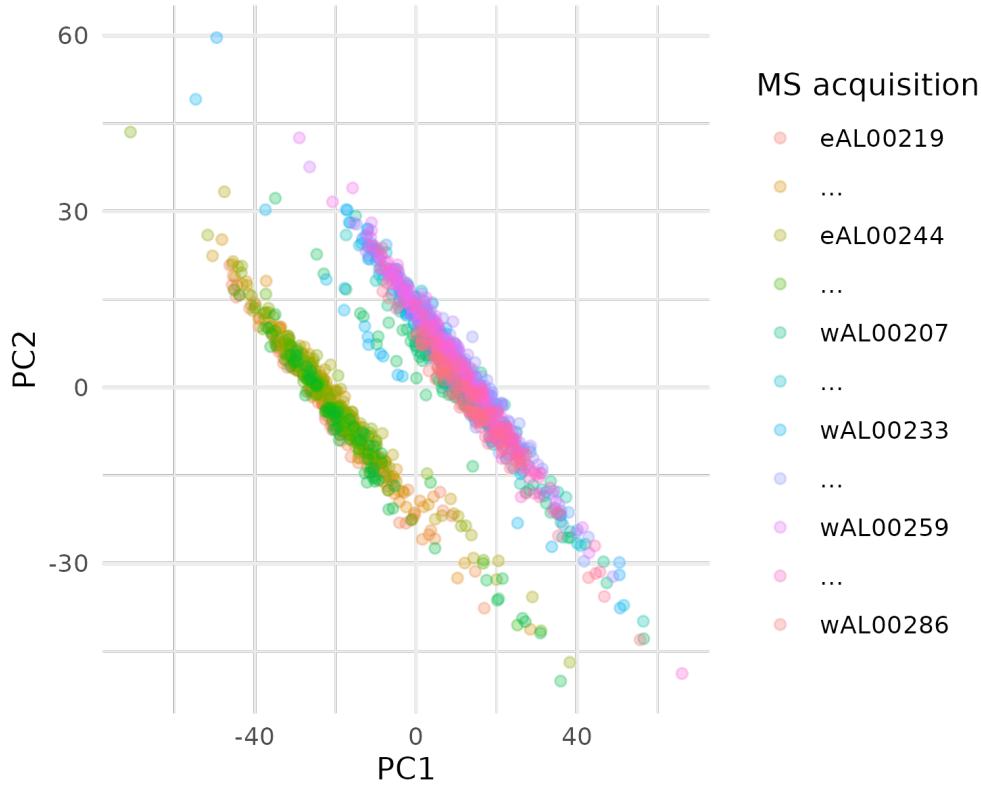
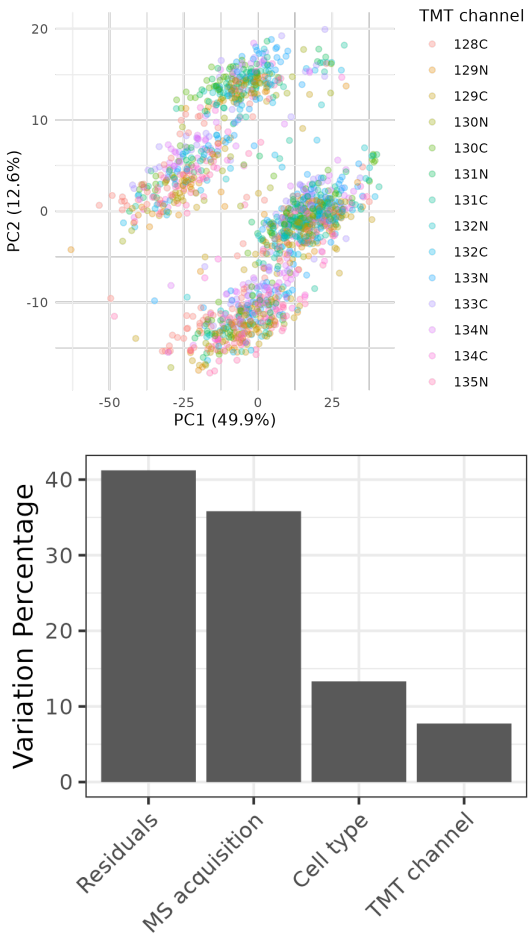


Figure: PCA on the **MS acquisition** effect matrix.

PCA on effect matrices

$$y = MS\ acquisition + TMT\ channel + Cell\ type + \epsilon$$



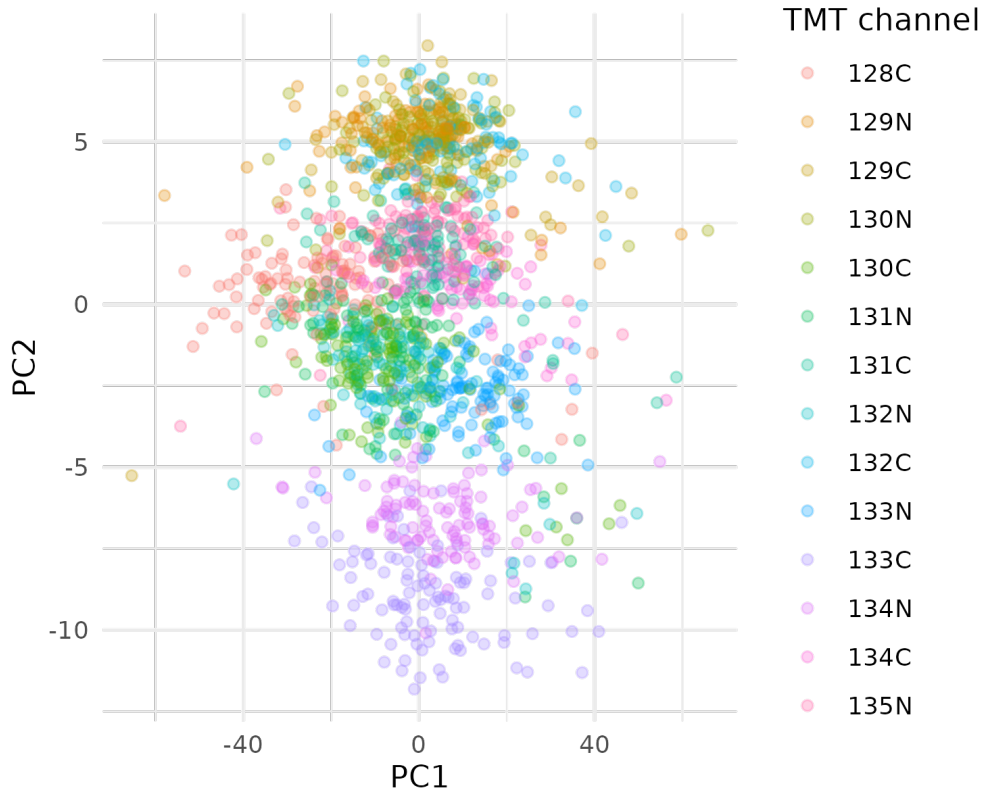
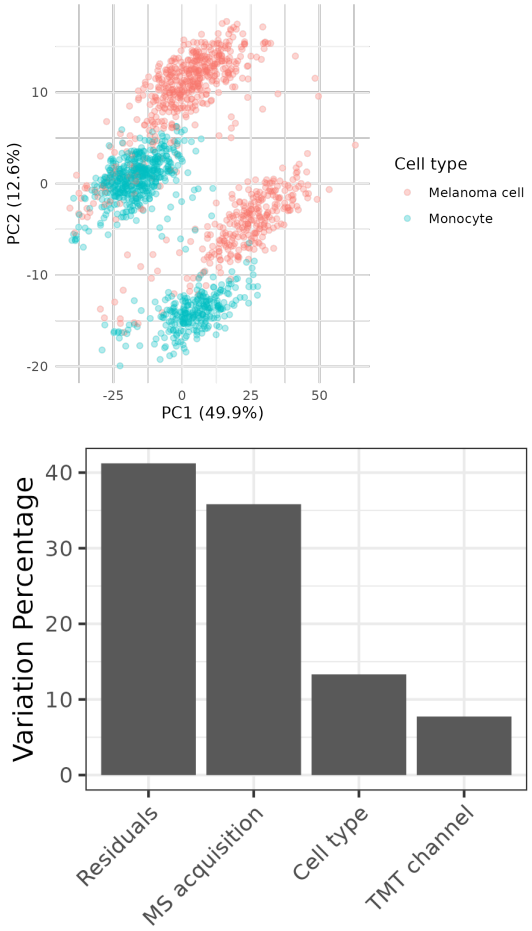


Figure: PCA on the **TMT channel** effect matrix.

PCA on effect matrices

$$y = MS\ acquisition + TMT\ channel + Cell\ type + \epsilon$$



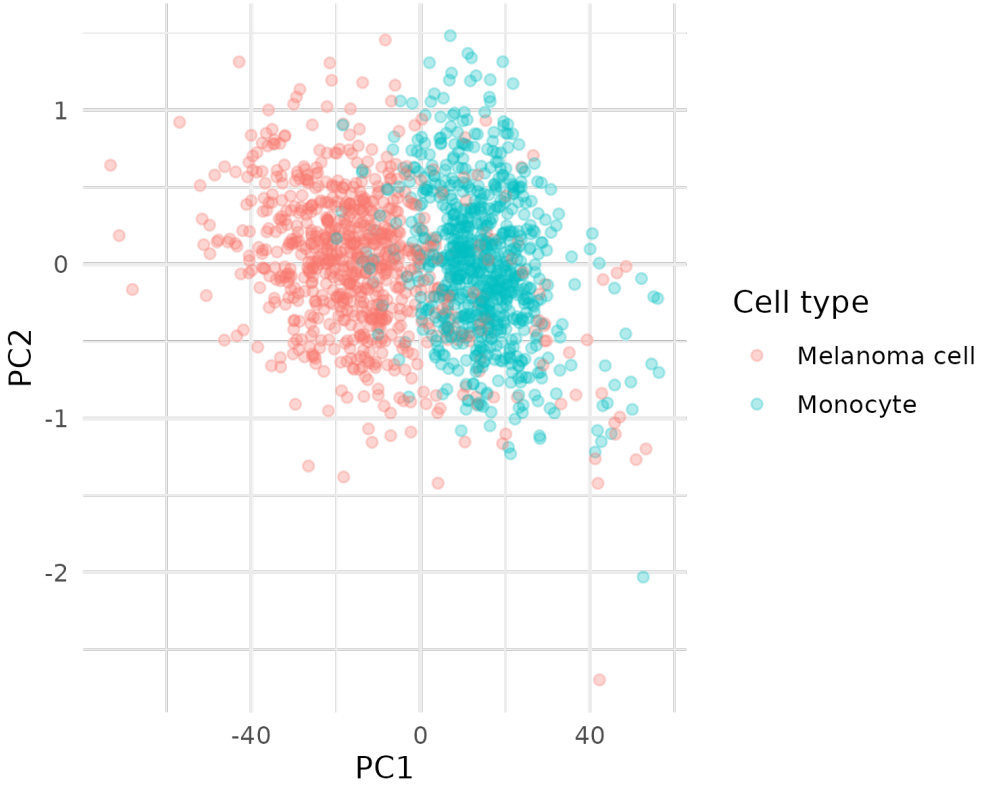
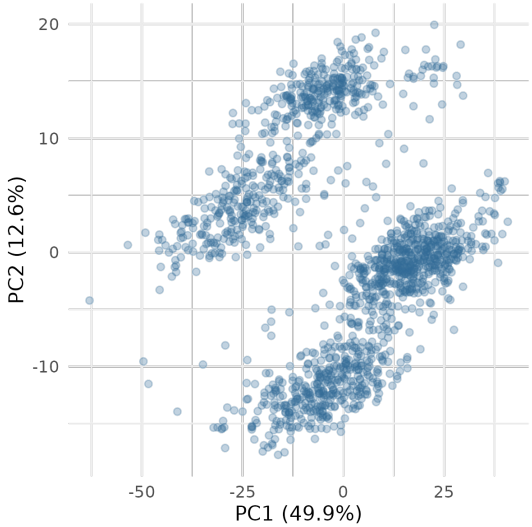
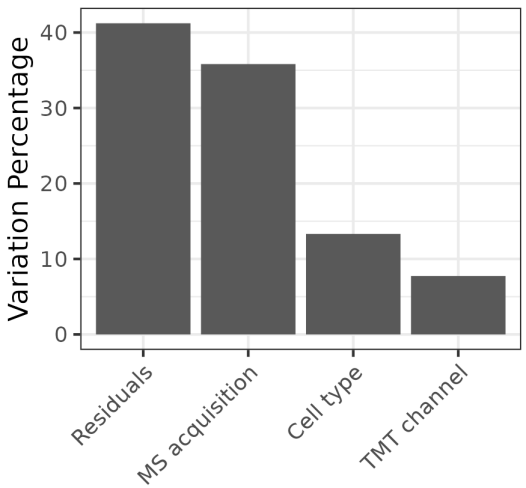


Figure: PCA on the **Cell type** effect matrix.

PCA on effect matrices

$$y = MS\ acquisition + TMT\ channel + Cell\ type + \epsilon$$





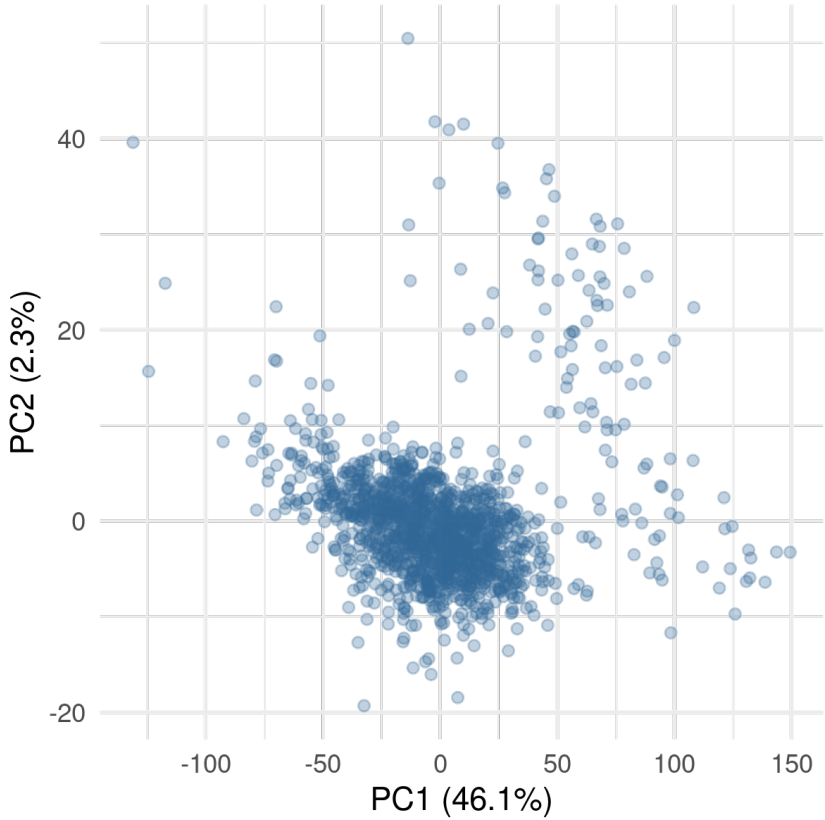
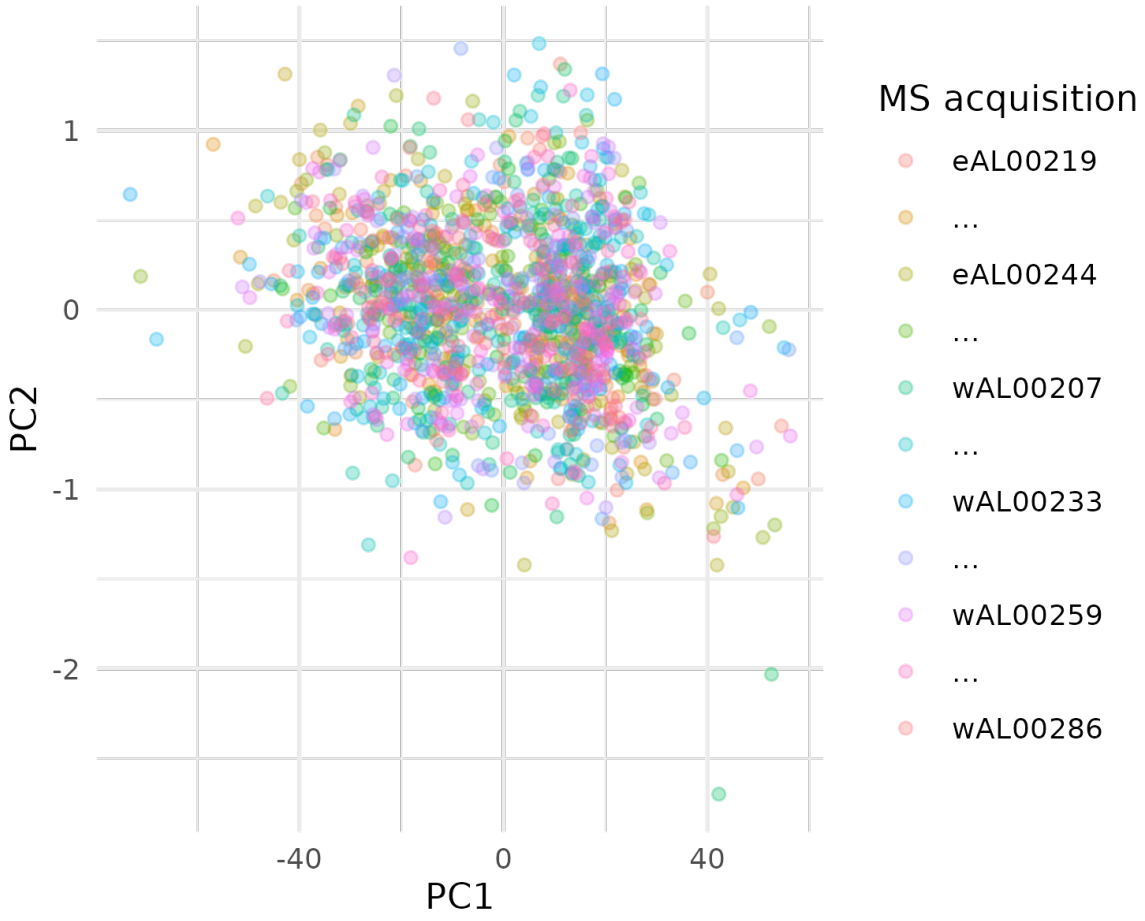


Figure: PCA on the **residuals** effect matrix.

Does it work: negative control

Do we have any MS acquitison batch *leftovers* in the cell type effect?

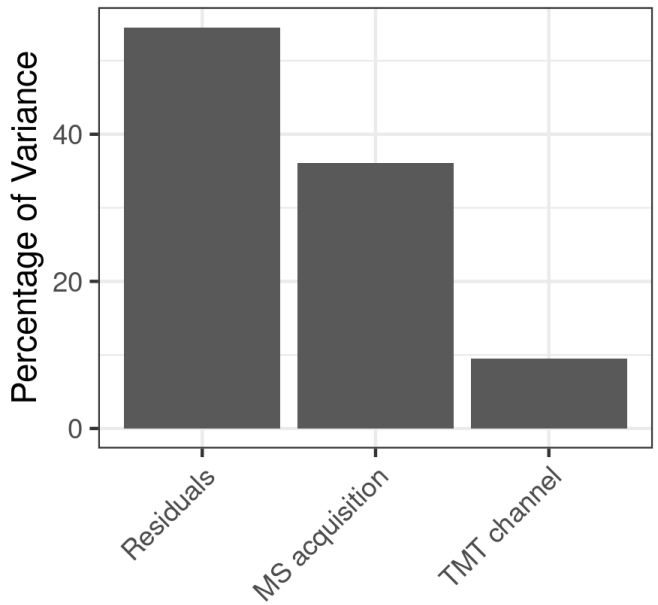


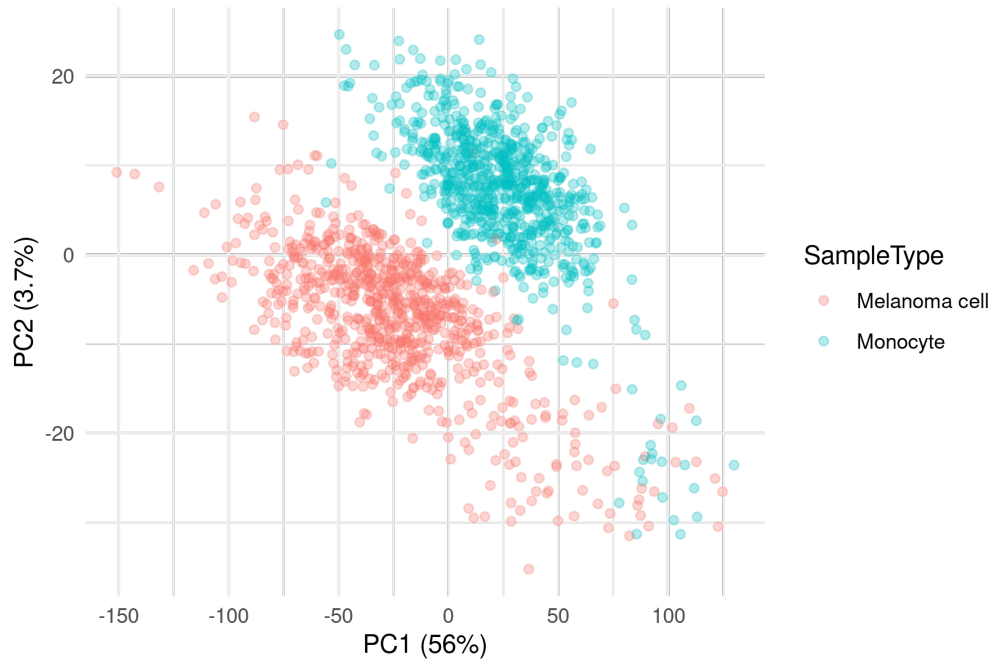
Does it work: positive control

$$y = MS\ acquisition + TMT\ channel + \epsilon$$

Does it work: positive control

$$y = MS\ acquisition + TMT\ channel + \epsilon$$





Does it work: new biology in the residuals

$$y = MS\ acquisition + TMT\ channel + Cell\ type + \epsilon$$

Does it work: new biology in the residuals

$$y = MS\ acquisition + TMT\ channel + Cell\ type + \epsilon$$

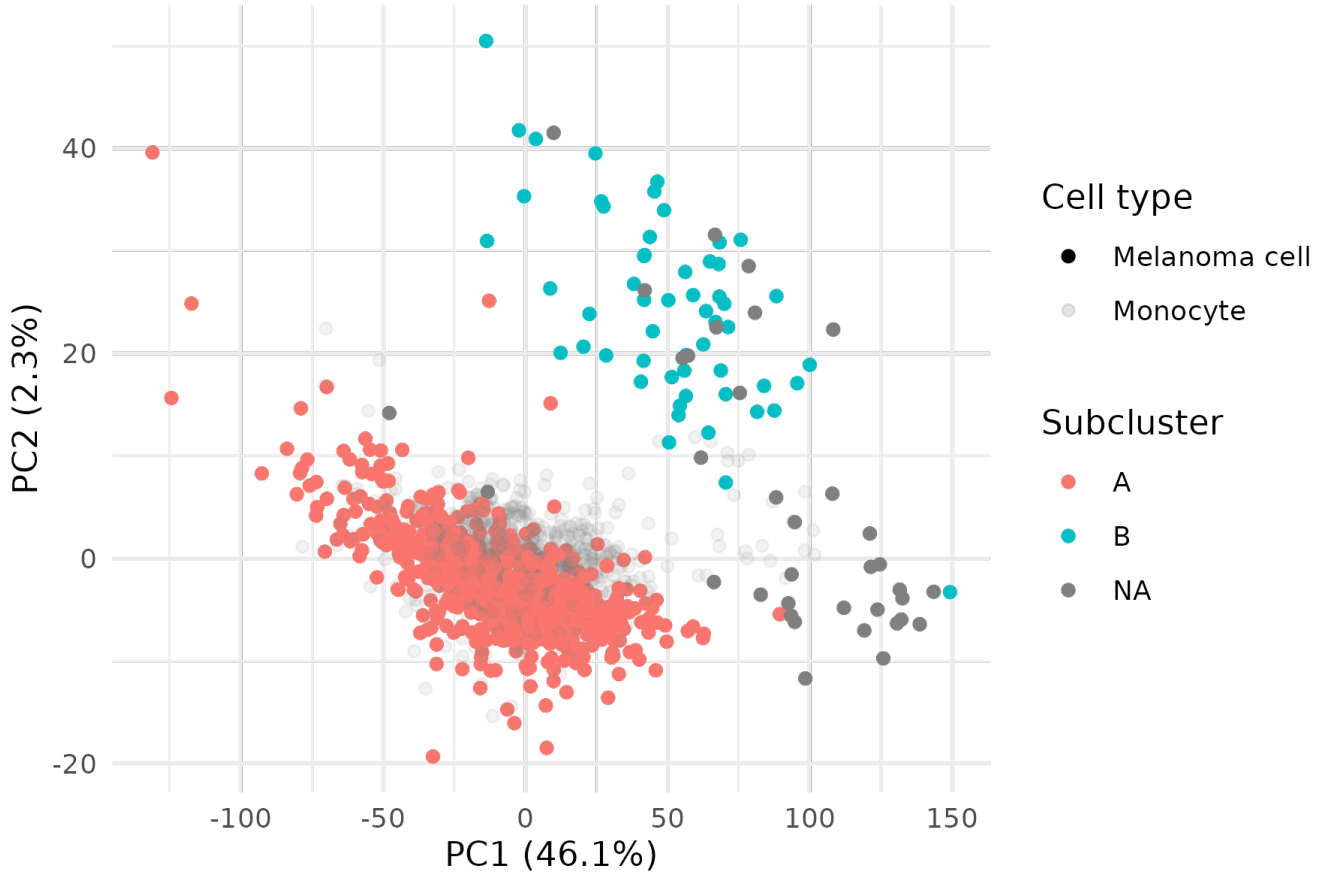


Figure: **Melanoma subpopulations**: transcriptomic signature associated with a cell state that is more likely to resist treatment by the cancer drug vemurafenib (clusters A and B from Leduc et al. (2022)).

Outline

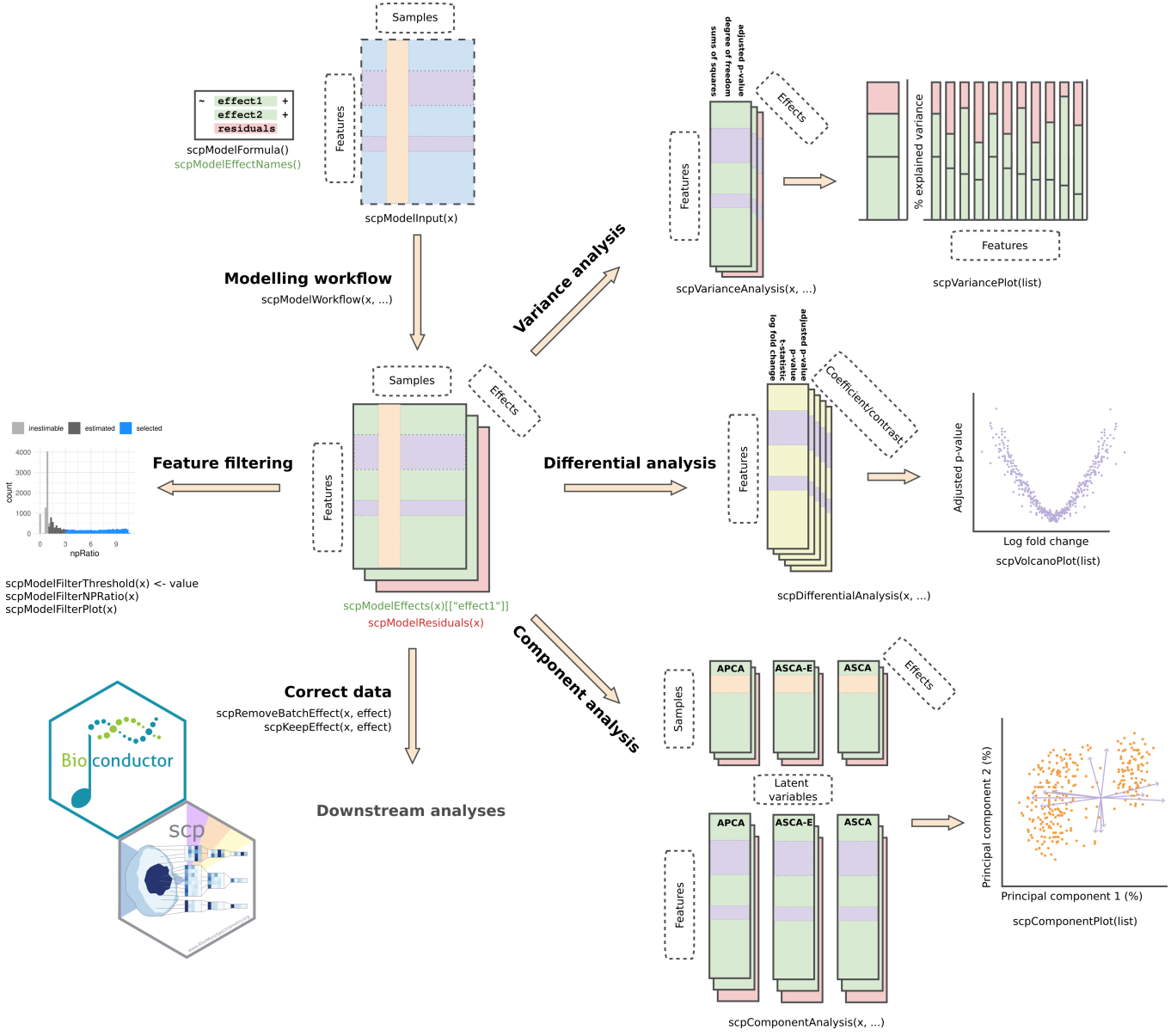


Figure: scp package - scplainer: using linear models to understand mass spectrometry-based single-cell proteomics data (Vanderaa and Gatto, 2024).

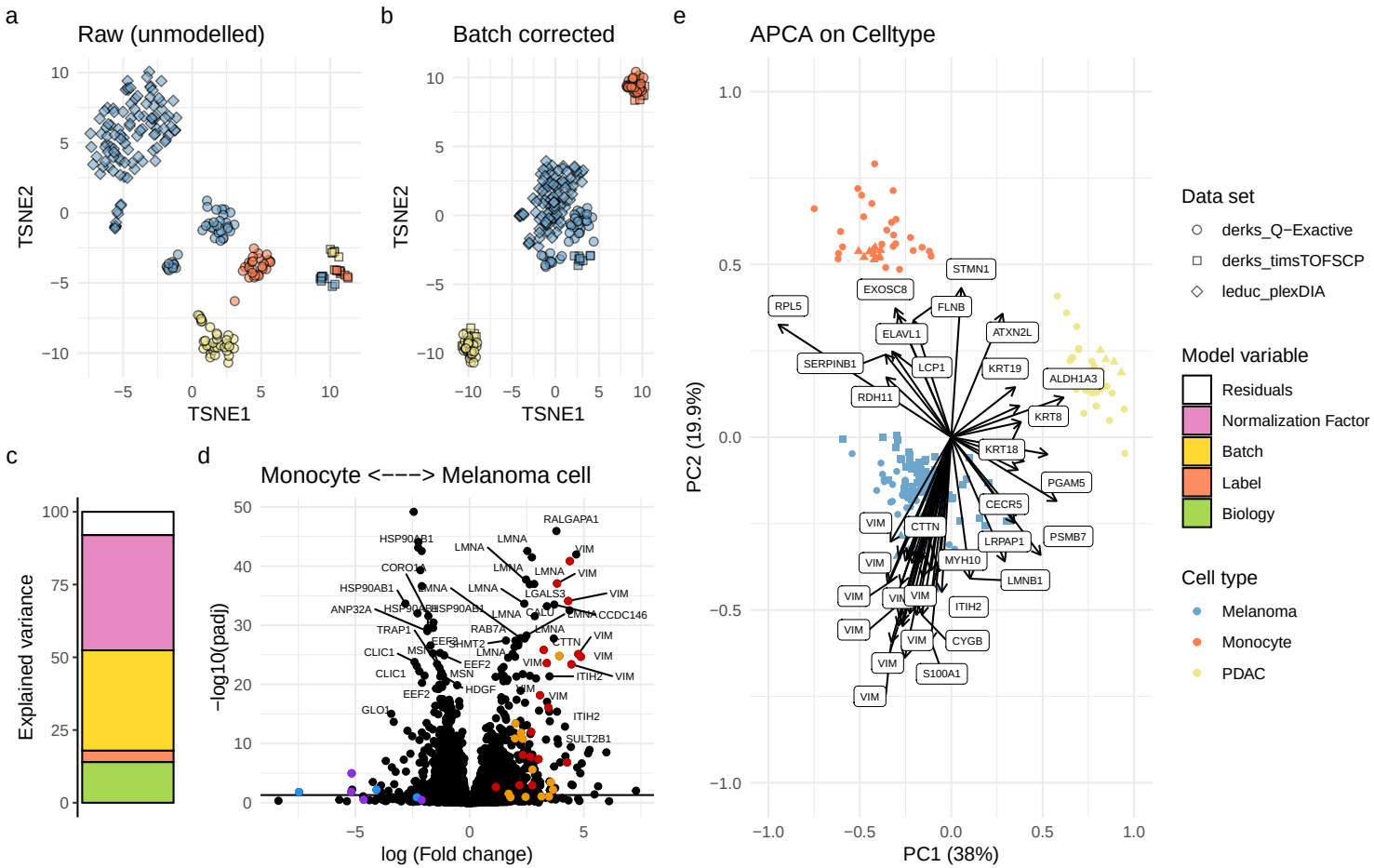
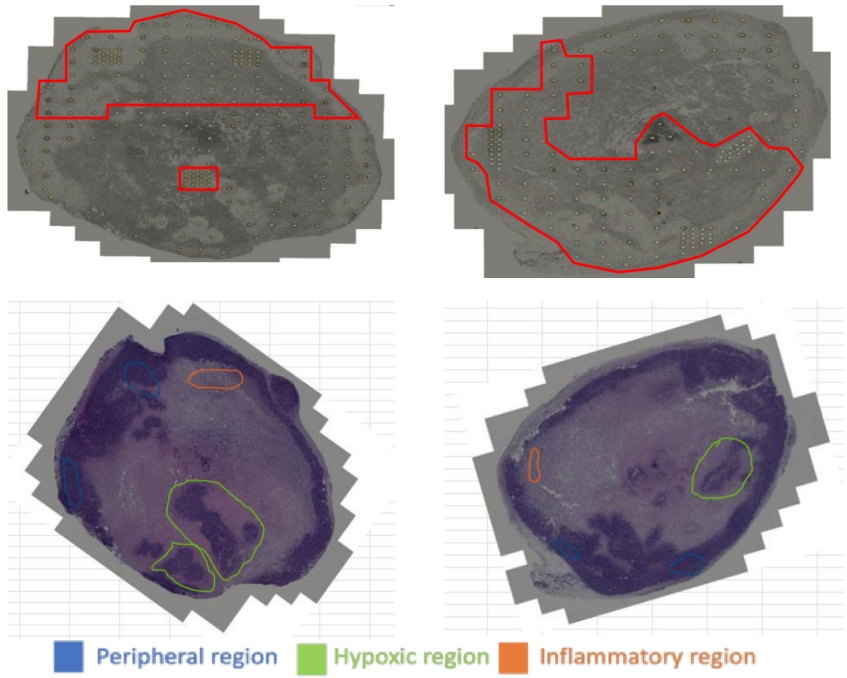


Figure: scplainer – variance, differential and component analysis, integration

Outline

Spatial proteomics



Xenograft of human breast tumor, control and treated. About 120 laser capture microdissected 2 mm spots per section. DIA MS, 1500 proteins/spot (G. Mazzucchelli, Liège). Analysed using existing and standardised infrastructure for (single-cell) quantitative proteomics, and spatial omics (`SpatialExperiment` Righelli et al. (2022)).

Spatial and single-cell proteomics

- Spatial proteomics: Automated micro-dissection along a high resolution grid (25 μm).
- Acquire single-cell proteomics (from the same biopsy) to define a sample-specific reference.
- $\rightarrow$ deconvolution of individual spots

Spatial and single-cell proteomics

- Spatial proteomics: Automated micro-dissection along a high resolution grid (25 μm).
- Acquire single-cell proteomics (from the same biopsy) to define a sample-specific reference.
- $\rightarrow$ deconvolution of individual spots

Other spatial proteomics references

- Mund et al. (2022) *Deep Visual Proteomics defines single-cell identity and heterogeneity*.
- Davis et al. (2023) *Deep topographic proteomics of a human brain tumour*.

Outline

Conclusions

- Many experimental and computational workflows. Different workflows $\rightarrow$ different results.
- We need a flexible and **principled computational approach** $\rightarrow$ control what we do, to guarantee the validity or our results.
- **Residuals** – what we don't know (yet), generally what we are most interested in.
- Showed component analysis, differential abundance, analysis of variance. Also clustering, trajectory analysis, ... based on the batch-corrected/normalised effect matrices.

Conclusions

- Many experimental and computational workflows. Different workflows $\rightarrow$ different results.
- We need a flexible and **principled computational approach** $\rightarrow$ control what we do, to guarantee the validity or our results.
- **Residuals** – what we don't know (yet), generally what we are most interested in.
- Showed component analysis, differential abundance, analysis of variance. Also clustering, trajectory analysis, ... based on the batch-corrected/normalised effect matrices.

- Work openly and reproducibly! (Markowetz, 2015)
- Importance of the **experimental design** (Gatto et al., 2023).

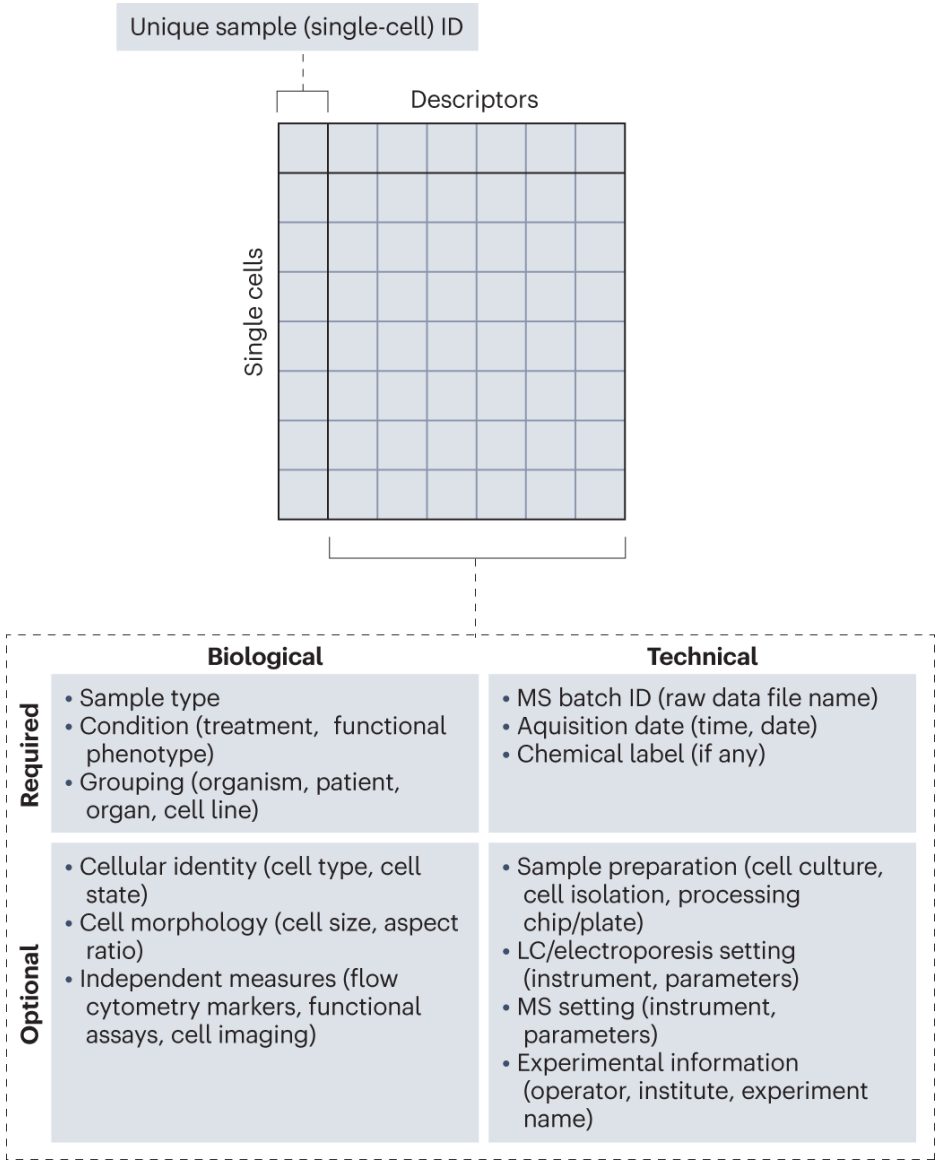


Figure: Initial recommendations for performing, benchmarking and reporting single-cell proteomics experiments. Suggested descriptors of single-cell proteomic samples (Gatto et al., 2023).

References I

Simon Davis, Connor Scott, Janina Oetjen, Philip D Charles, Benedikt M Kessler, Olaf Ansorge, and Roman Fischer. Deep topographic proteomics of a human brain tumour. *Nat. Commun.*, 14(1):7710, November 2023.

Laurent Gatto, Ruedi Aebersold, Juergen Cox, Vadim Demichev, Jason Derks, Edward Emmott, Alexander M Franks, Alexander R Ivanov, Ryan T Kelly, Luke Khoury, Andrew Leduc, Michael J MacCoss, Peter Nemes, David H Perlman, Aleksandra A Petelski, Christopher M Rose, Erwin M Schoof, Jennifer Van Eyk, Christophe Vanderaa, John R Yates, and Nikolai Slavov. Initial recommendations for performing, benchmarking and reporting single-cell proteomics experiments. *Nat. Methods*, pages 1–12, March 2023.

Mo Hu, Yutong Zhang, Yuan Yuan, Wenping Ma, Yinghui Zheng, Qingqing Gu, and X Sunney Xie. Correlated protein modules revealing functional coordination of interacting proteins are detected by Single-Cell proteomics. *J. Phys. Chem. B*, 127 (27):6006 – 6014, July 2023.

Andrew Leduc, R Gray Huffman, Joshua Cantlon, Saad Khan, and Nikolai Slavov. Exploring functional protein covariation across single cells using nPOP. *Genome Biol.*, 23(1):1–31, December 2022.

Florian Markowetz. Five selfish reasons to work reproducibly. *Genome Biol.*, 16:274, December 2015.

References II

Andreas Mund, Fabian Coscia, András Kriston, Réka Hollandi, Ferenc Kovács, Andreas-David Brunner, Ede Migh, Lisa Schweizer, Alberto Santos, Michael Bzorek, Soraya Naimy, Lise Mette Rahbek-Gjerdrum, Beatrice Dyring-Andersen, Jutta Bulkescher, Claudia Lukas, Mark Adam Eckert, Ernst Lengyel, Christian Gnann, Emma Lundberg, Peter Horvath, and Matthias Mann. Deep visual proteomics defines single-cell identity and heterogeneity. *Nat. Biotechnol.*, May 2022.

Dario Righelli, Lukas M Weber, Helena L Crowell, Brenda Pardo, Leonardo Collado-Torres, Shila Ghazanfar, Aaron T L Lun, Stephanie C Hicks, and Davide Risso. SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using bioconductor. *Bioinformatics*, 38(11):3128–3131, May 2022.

Nikolai Slavov. Learning from natural variation across the proteomes of single cells. *PLoS Biol.*, 20(1):e3001512, January 2022.

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.

Michel Thiel, Baptiste Féraud, and Bernadette Govaerts. ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. *J. Chemom.*, 31(6):e2895, June 2017.

References III

Christophe Vanderaa and Laurent Gatto. Replication of single-cell proteomics data reveals important computational challenges. *Expert Rev. Proteomics*, October 2021.

Christophe Vanderaa and Laurent Gatto. The current state of Single-Cell proteomics data analysis. *Curr Protoc*, 3(1):e658, January 2023a.

Christophe Vanderaa and Laurent Gatto. Revisiting the thorny issue of missing values in single-cell proteomics. *arXiv [q-bio.QM]*, April 2023b.

Christophe Vanderaa and Laurent Gatto. scplainer: using linear models to understand mass spectrometry-based single-cell proteomics data. *bioRxiv*, 2024. doi: 10.1101/2023.12.14.571792.

UCLouvain

Acknowledgments

- Dr Christophe Vanderaa (CBIO, now UGent)
- Dr Gabriel Mazzucchelli *et al.* (ULiège)
- Dr Luojiao Huang (MERLIN, Maastricht University), OpenMS fellow

- **Computational Biology and Bioinformatics**
  de Duve Institute
  UCLouvain
  lgatto.github.io/cbio-lab

Funding: Fonds National de la Recherche Scientifique - **FNRS**.