

# How to analyse single-cell proteomics data and focus on the underlying biology?

Laurent Gatto

de Duve Institute, UCLouvain

25 April 2025

# Outline

Single-cell proteomics: introduction

SCP data/analysis - round 1

Computational challenges

A principled approach to SCP data analysis - round 2

Implementation - `scp` and `scclaimer`

Conclusions

# Single-cell technologies unravel cellular heterogeneity

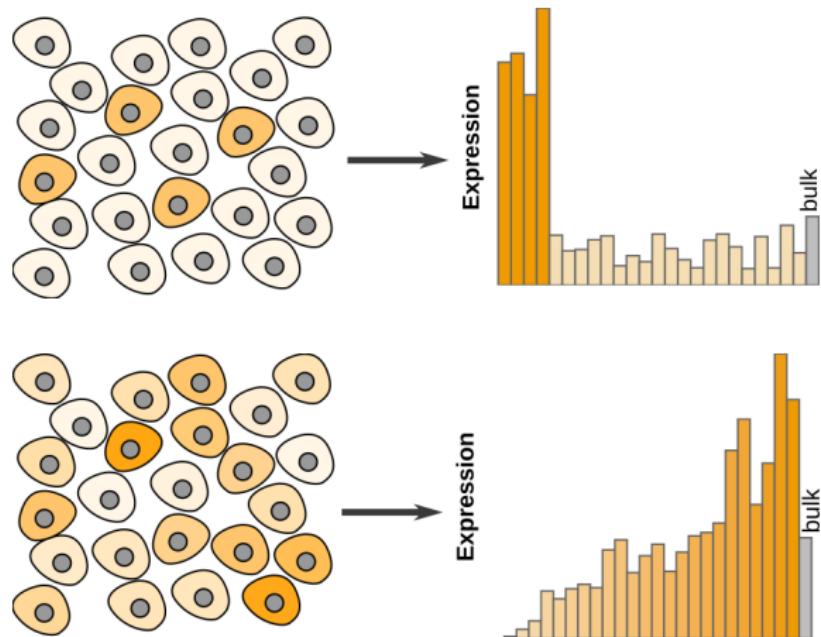
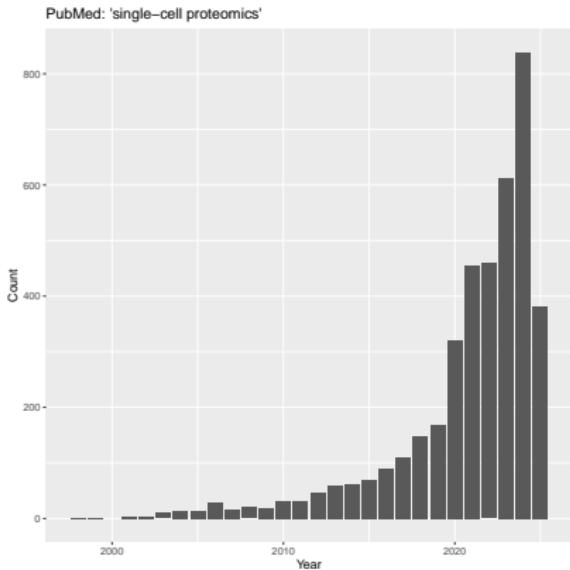


Figure: Cell types and cell states, subpopulation identification, differentiation trajectories (in the absence of known markers).



**August 2019:** in a [Nature Methods Technology Feature<sup>a</sup>](#), Vivien Marx *dreamt* of single-cell proteomics.

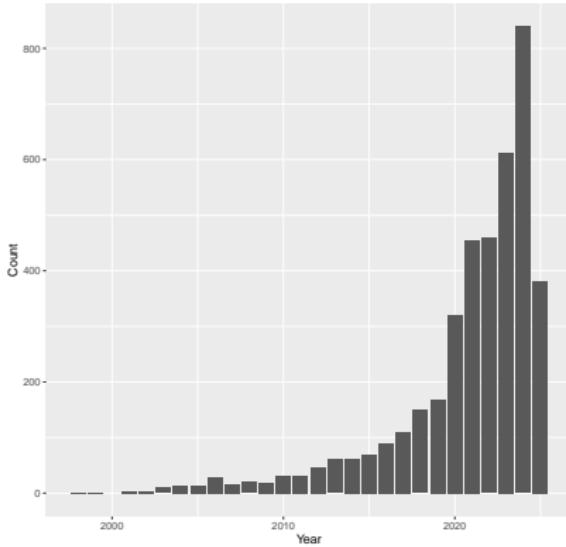
**March 2023:** Nature Methods published a special issue with a [Focus on single-cell proteomics<sup>b</sup>](#) and [Initial recommendations for performing, benchmarking and reporting single-cell proteomics experiments<sup>c</sup>](#).

<sup>a</sup> [10.1038/s41592-019-0540-6](https://doi.org/10.1038/s41592-019-0540-6)

<sup>b</sup> [www.nature.com/collections/bdfhafhdeb](https://www.nature.com/collections/bdfhafhdeb)

<sup>c</sup> [10.1038/s41592-023-01785-3](https://doi.org/10.1038/s41592-023-01785-3)

PubMed: 'single-cell proteomics'



**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	<b>functional</b>

- ▶ FC vs. SCP → unsupervised.
- ▶ RNA → intention vs. protein → action.
- ▶ Inference of direct regulatory interactions with minimal assumptions ([Slavov, 2022](#); [Hu et al., 2023](#)).
- ▶ Post-translational modifications.

# Outline

Single-cell proteomics: introduction

SCP data/analysis - round 1

Computational challenges

A principled approach to SCP data analysis - round 2

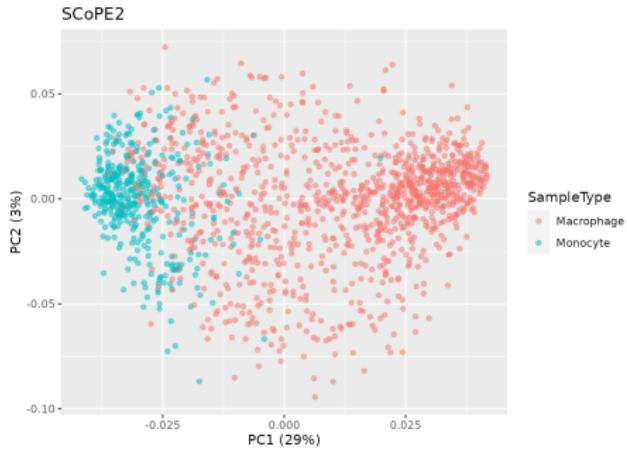
Implementation - `scp` and `scclaimer`

Conclusions

# 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

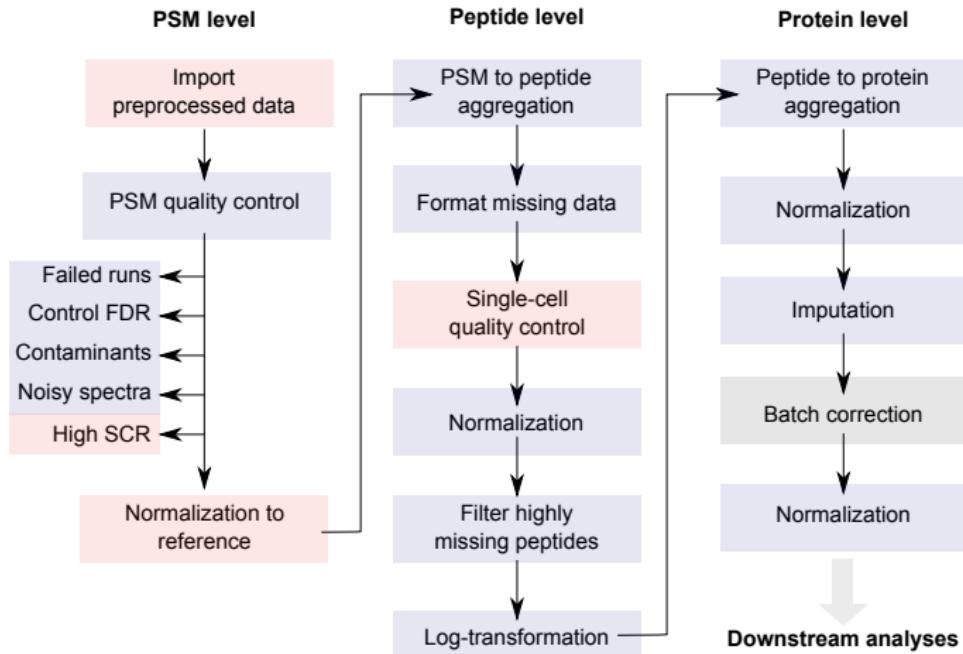


Figure: **Overview of the key steps performed in the SCoPE2 pipeline** (Vanderaa and Gatto, 2021). Blue boxes: QFeatures. Red boxes: scp. Gray box: sva::ComBat.

# Outline

Single-cell proteomics: introduction

SCP data/analysis - round 1

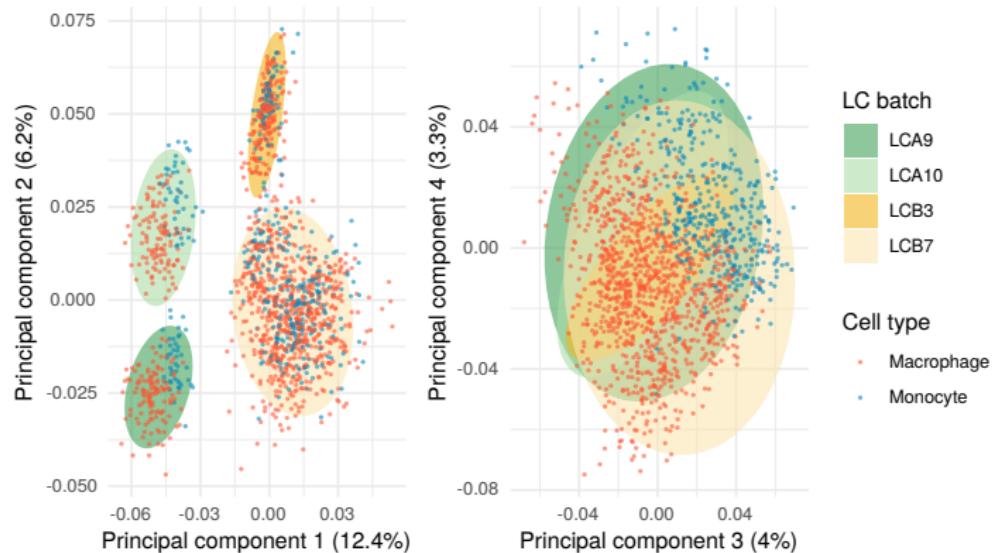
Computational challenges

A principled approach to SCP data analysis - round 2

Implementation - `scp` and `scclaimer`

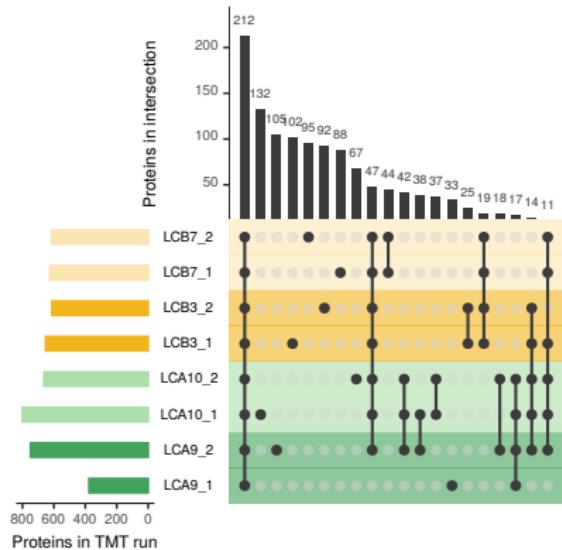
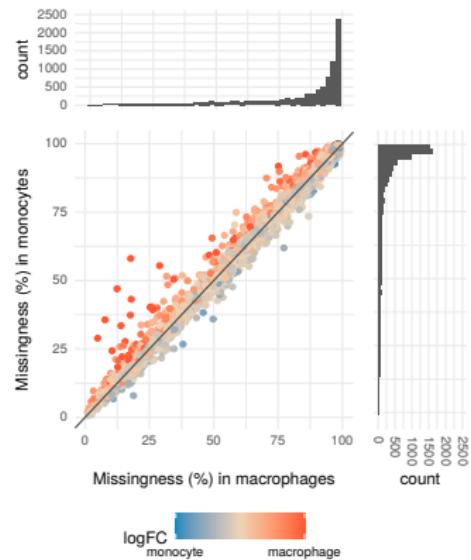
Conclusions

# 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

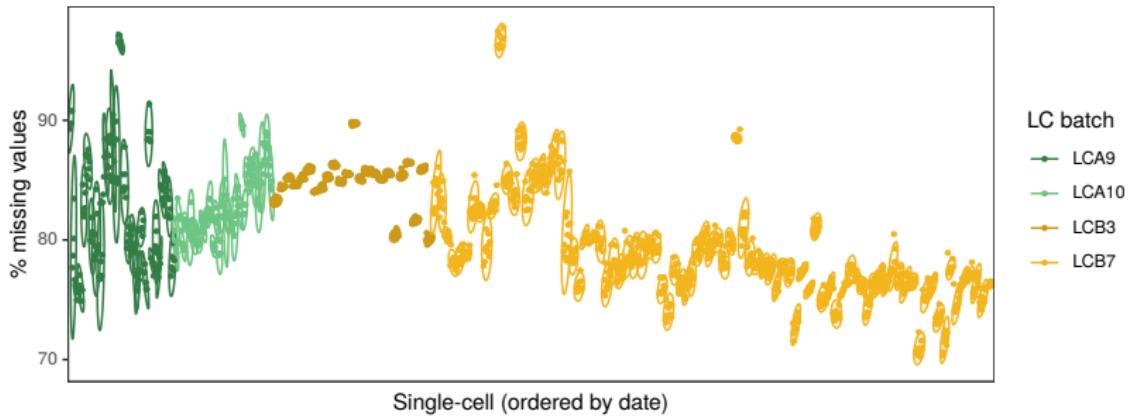


Figure: Influence of batch on data missingness ([Vanderaa and Gatto, 2021](#)).

## Data analyses review

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

# Replication

The screenshot shows a web application interface for 'SCP.replication' version 0.2.1. At the top, there are tabs for 'SCP.replication' (selected), '0.2.1', 'Home', 'Reference', and 'Articles'. A dropdown menu 'Articles' is open, showing a list of publications:

- Reproduction of the SCoPE2 analysis (Specht et al. 2021)
- Exploring the autoPOTS data (Liang et al. 2020)
- Reproduction of the AML model analysis (Schoof et al. 2021)
- Reproduction of the hair-cell development analysis (Zhu et al. 2019, eLife)

To the right of the list, there are several buttons with associated text:

- Filter the PSM data
- Normalize to reference
- Aggregate PSM data to peptide data
- Join the SCoPE2 sets in one assay
- Filter single-cells based on median CV
- Process the peptide data
- Aggregate peptide data to protein data
- Process the protein data
- Benchmarking the replication
- Conclusion
- Requirements
- Reference

**Abstract**

Recent advances in sample preparation, processing and mass spectrometry (MS) have allowed the emergence of MS-based single-cell proteomics (SCP). This vignette presents a robust and standardized workflow to reproduce the data analysis of SCoPE2, one of the pioneering works in the field developed by the Slavov Lab. The implementation uses well-defined Bioconductor classes that provide powerful tools for single-cell RNA sequencing and for shotgun proteomics. We demonstrate that our pipeline can reproduce the SCoPE2 analysis using only a few lines of code.

**Introduction**

SCoPE2 (Specht et al. (2021)) is the first mass spectrometry (MS)-based single cell proteomics (SCP) protocol that has been used to profile thousands of proteins in thousands of single-cells. This is a technical milestone for

<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

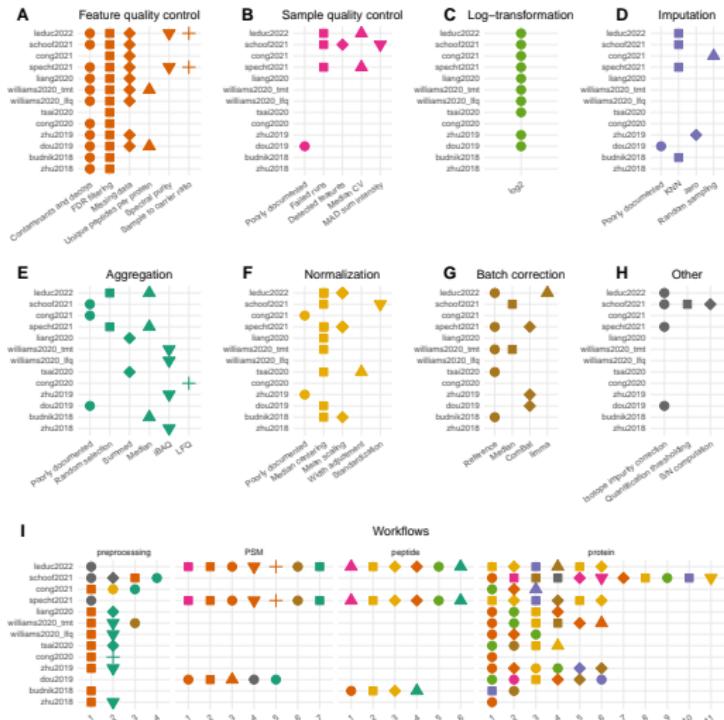


Figure: Single-cell data processing: **one workflow per paper/lab.** (Vanderaa and Gatto, 2023a)

## 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

Single-cell proteomics: introduction

SCP data/analysis - round 1

Computational challenges

A principled approach to SCP data analysis - round 2

Implementation - `scp` and `scclaimer`

Conclusions

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](#)), implemented as the [sciplainer](#) approach ([Vanderaa and Gatto, 2024](#)) in the [scp](#) package ([Vanderaa and Gatto, 2023a](#)).

## (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

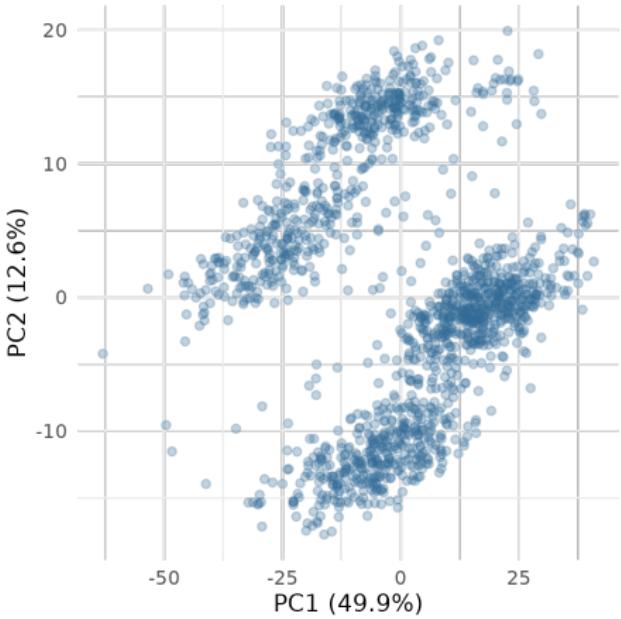
## (3) Principal Component Analysis

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**



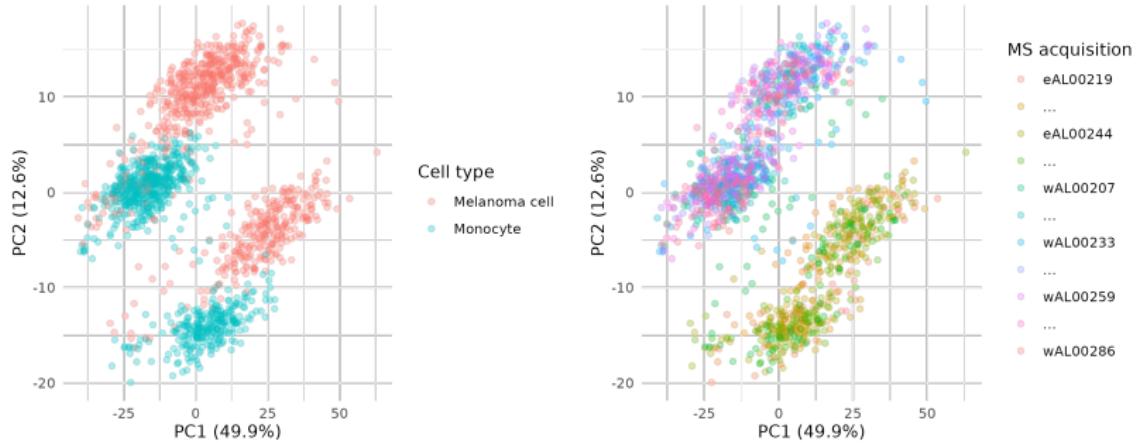


Figure: Melanoma cells and monocytes (left) acquired across multiple acquisition batches (right) (Leduc et al., 2022).

$$y = \textcolor{blue}{MS \ acquisition} + \textcolor{blue}{TMT \ channel} + \textcolor{orange}{Cell \ type} + \epsilon$$

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

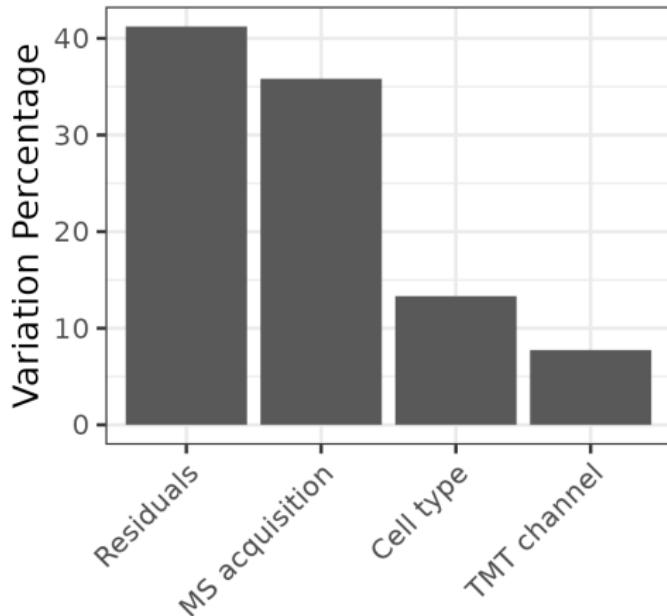


Figure: We are now in a position to **quantify known and unknown**

**effects:** percentages of explained variances of our explained (known) and unexplained (residuals) effects. NB: low biological variance  $\neq$  low quality!

# PCA on effect matrices

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

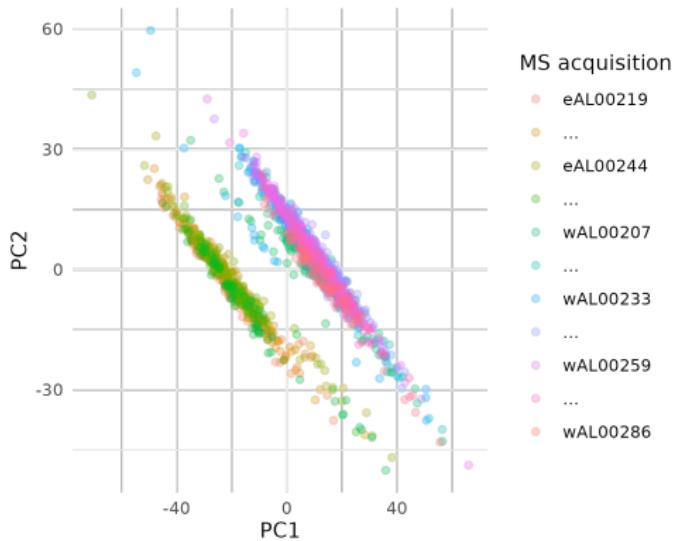
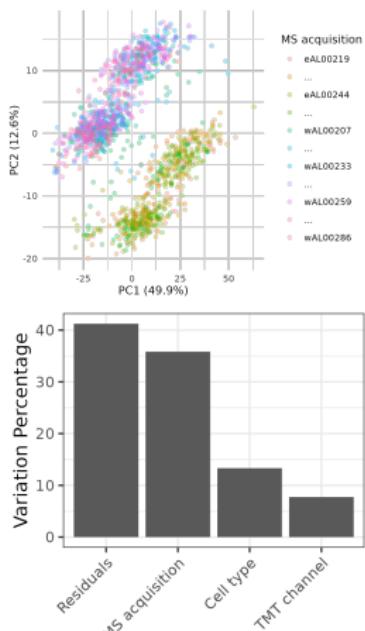


Figure: PCA on the **MS acquisition** effect matrix.

# PCA on effect matrices

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

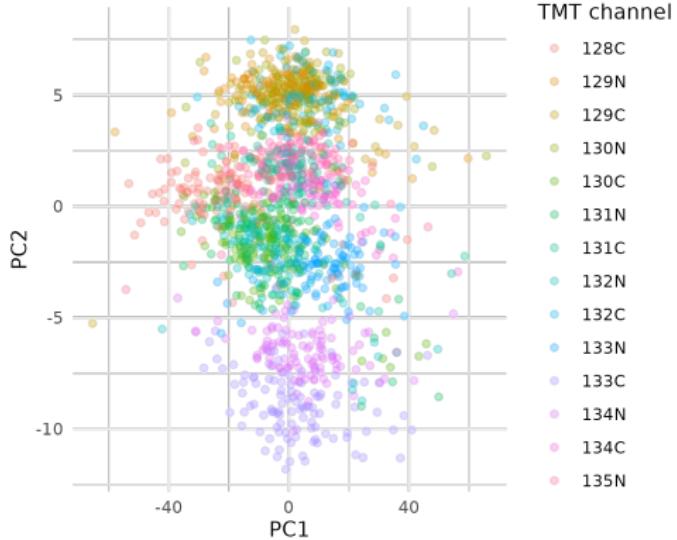
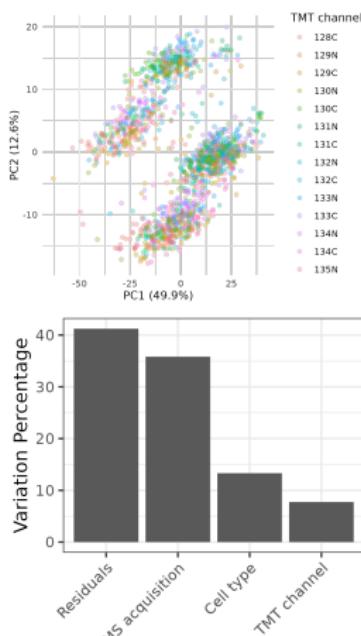


Figure: PCA on the **TMT channel** effect matrix.

# PCA on effect matrices

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

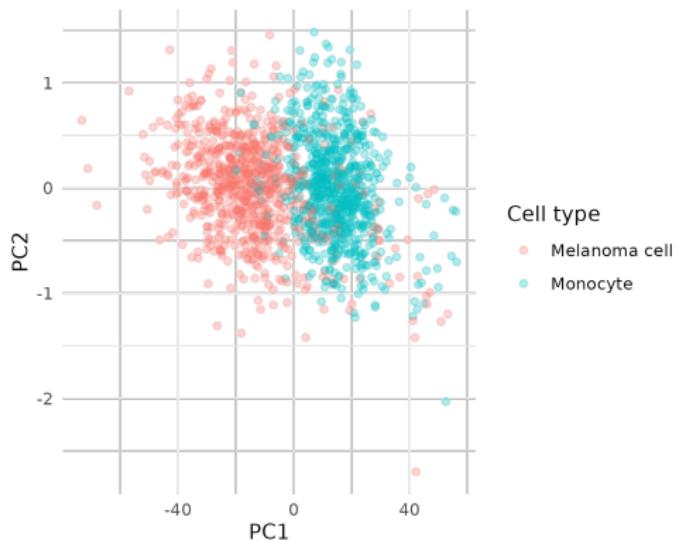
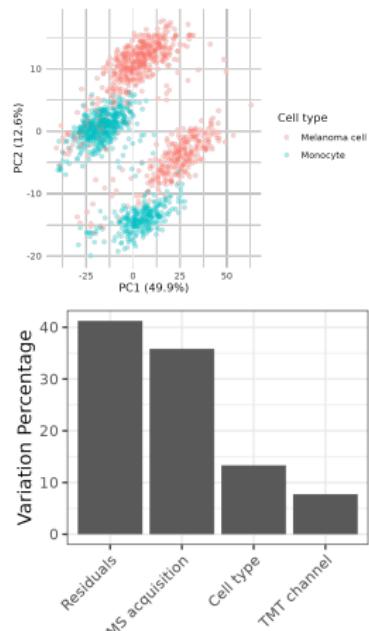


Figure: PCA on the **Cell type** effect matrix.

# PCA on effect matrices

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

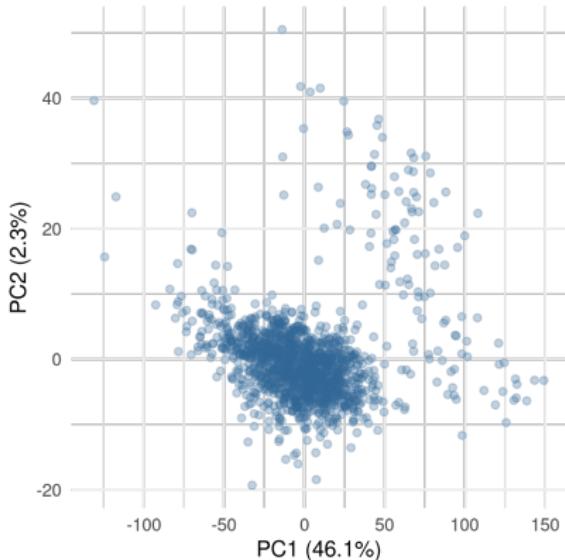
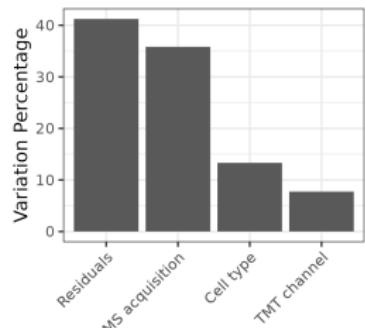
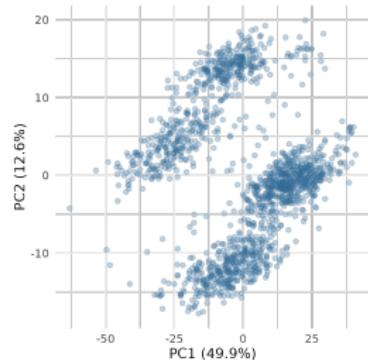
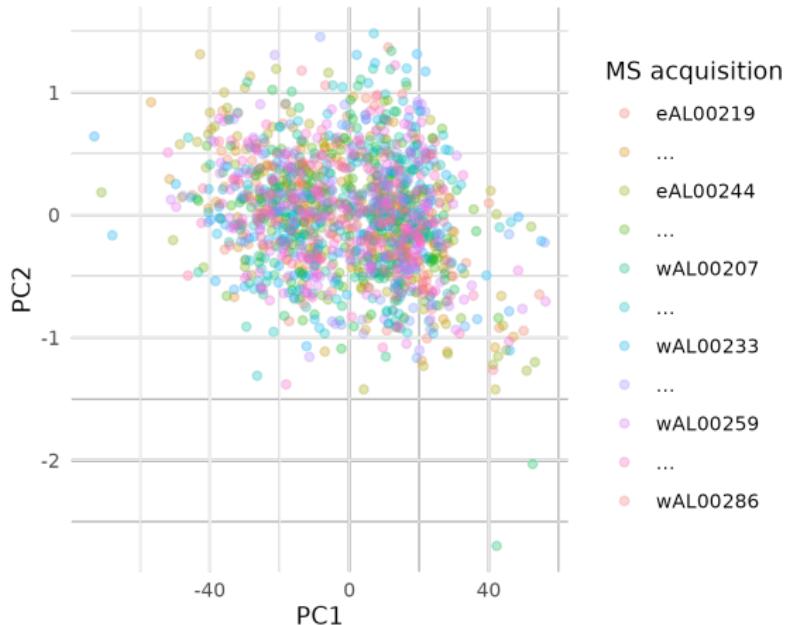


Figure: PCA on the **residuals** effect matrix.

# Does it work: negative control

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

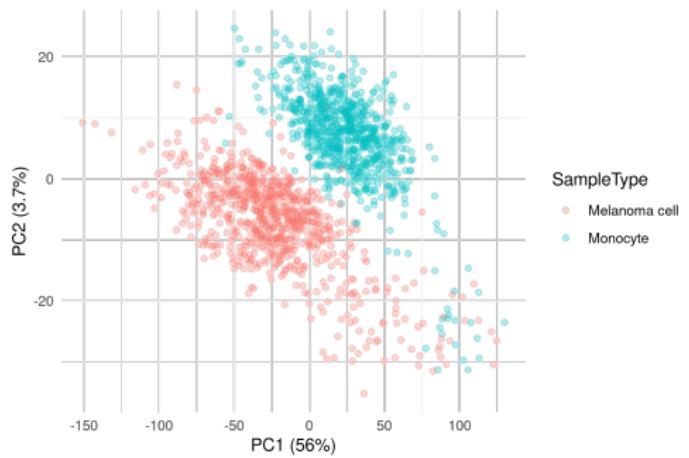
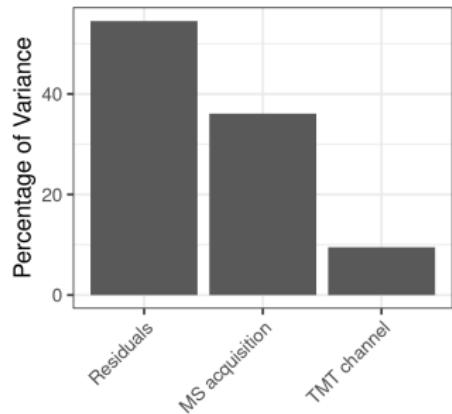


## Does it work: positive control

$$y = \text{MS acquisition} + \text{TMT channel} + \epsilon$$

# Does it work: positive control

$$y = \text{MS acquisition} + \text{TMT channel} + \epsilon$$

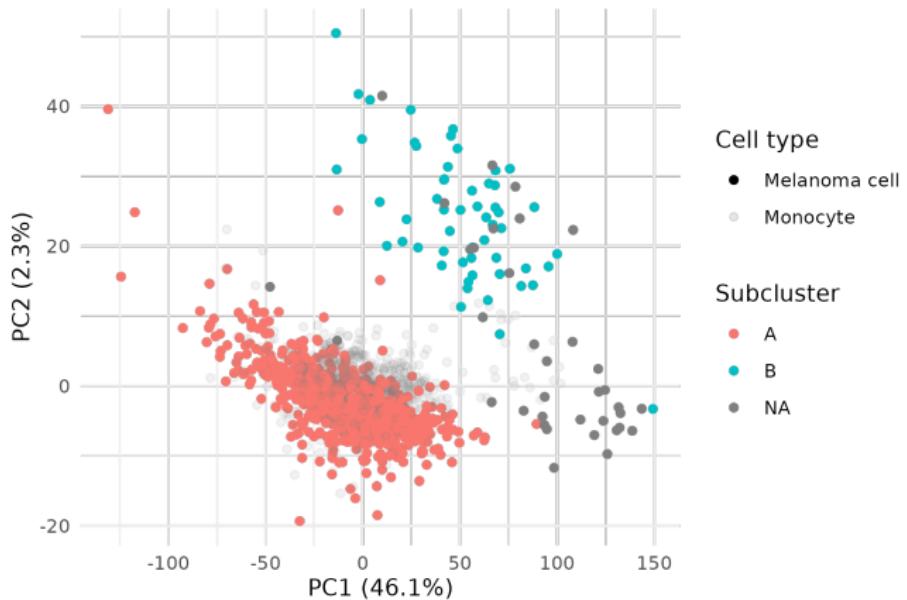


## Does it work: new biology in the residuals

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

# Does it work: new biology in the residuals

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



**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

Single-cell proteomics: introduction

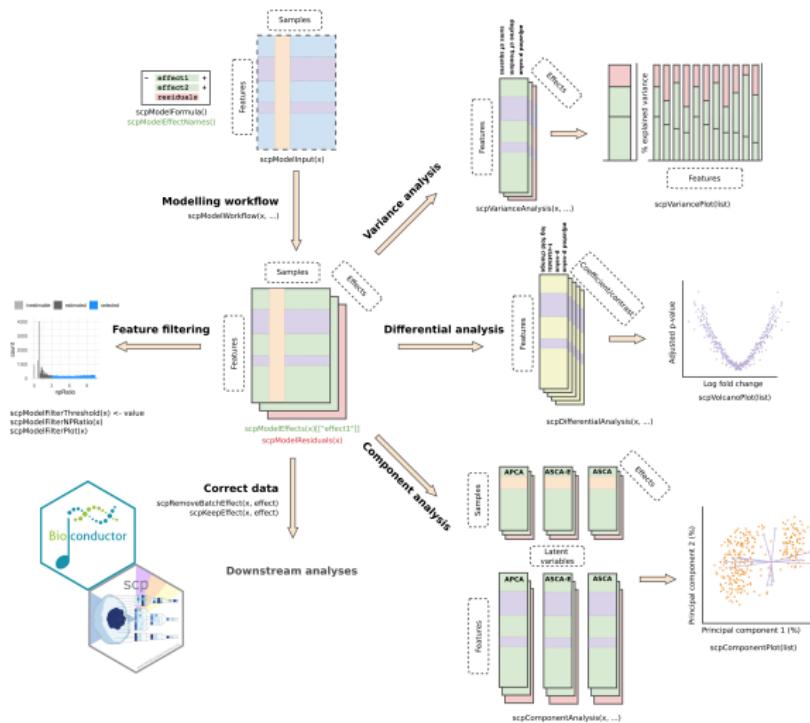
SCP data/analysis - round 1

Computational challenges

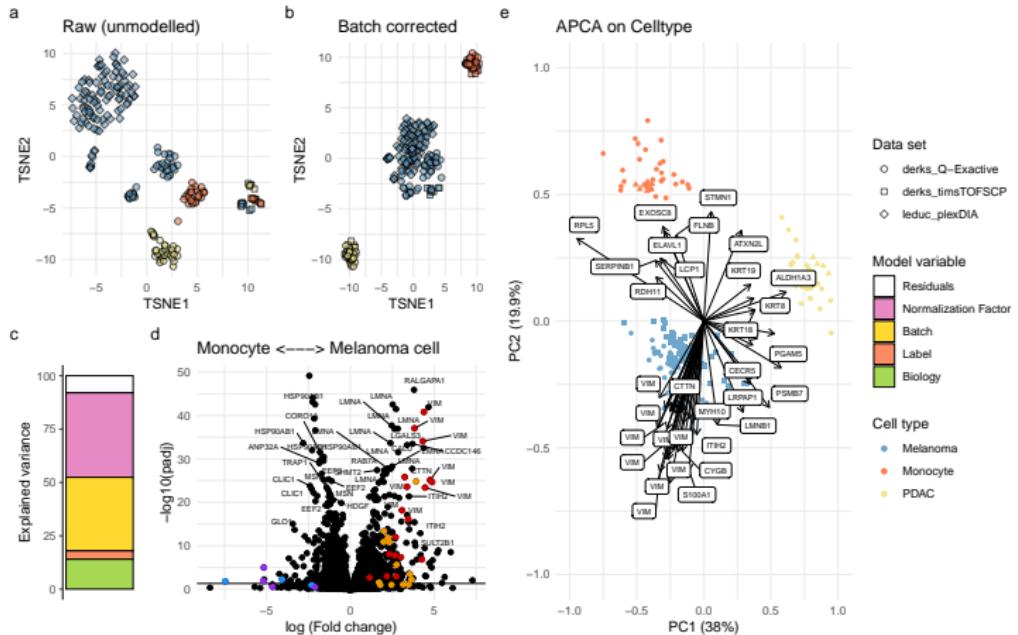
A principled approach to SCP data analysis - round 2

Implementation - `scp` and `scplainer`

Conclusions



**Figure:** `scp` package - **scplainer**: using linear models to understand mass spectrometry-based single-cell proteomics data ([Vanderaa and Gatto, 2024](#)). Part of and integrates with **Bioconductor** ([Huber et al., 2015](#)) tools.



**Figure:** scplainer – variance (c), differential (d) and component (e) analysis, integration (a, b)

# Outline

Single-cell proteomics: introduction

SCP data/analysis - round 1

Computational challenges

A principled approach to SCP data analysis - round 2

Implementation - `scp` and `scclaimer`

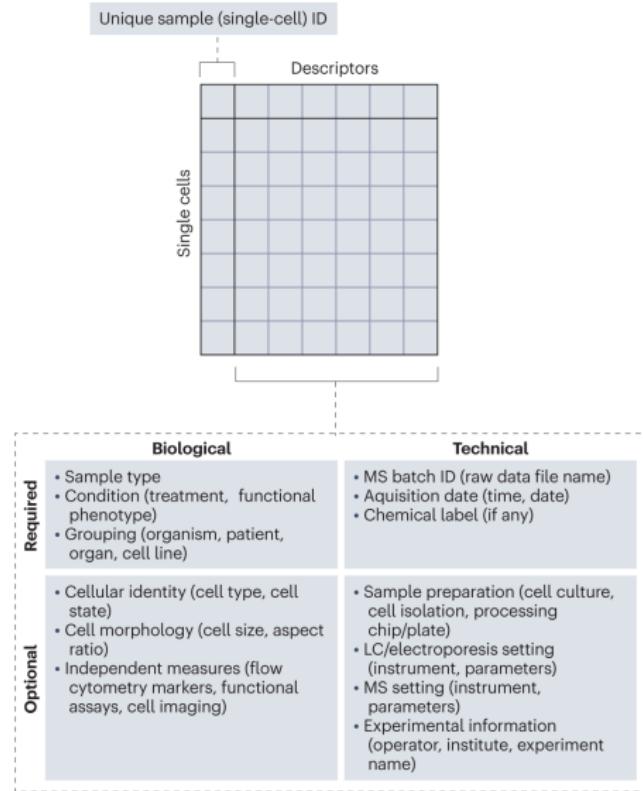
Conclusions

# Conclusions

- ▶ Many experimental and computational workflows. Different workflows → different results.
- ▶ We need a flexible and **principled computational approach** → control what we do, to guarantee the validity of 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.  
**Bioconductor tool kit.**

# Conclusions

- ▶ Many experimental and computational workflows. Different workflows → different results.
- ▶ We need a flexible and **principled computational approach** → control what we do, to guarantee the validity of 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.  
**Bioconductor tool kit.**
- ▶ 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 |

- Laurent Gatto, Ruedi Aebersold, Juergen Cox, Vadim Demichev, Jason Derks, Edward Emmott, Alexander M Franks, Alexander R Ivanov, Ryan T Kelly, Luke Khouri, 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.
- W Huber, V J Carey, R Gentleman, S Anders, M Carlson, B S Carvalho, H C Bravo, S Davis, L Gatto, T Girke, R Gottardo, F Hahne, K D Hansen, R A Irizarry, M Lawrence, M I Love, J MacDonald, V Obenchain, A K Oleś, H Pagès, A Reyes, P Shannon, G K Smyth, D Tenenbaum, L Waldron, and M Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*, 12(2):115–21, Jan 2015. doi: 10.1038/nmeth.3252.
- 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

- 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.
- 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.

## Acknowledgments

- ▶ **Computational Biology and Bioinformatics**, de Duve Institute, UCLouvain – [lgatto.github.io/cbio-lab](https://lgatto.github.io/cbio-lab)
- ▶ Dr Christophe Vanderaa (CBIO, now UGent)
- ▶ Prof Nikolai Slavov (Northeastern University)

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