How to analyse single-cell proteomics data and focus on the underlying biology?
Your results are only as good as your method and software

Laurent Gatto

de Duve Institute, UCLouvain

25 September 2025

Computational Systems Biology of Cancer, Paris, 25 Sept 2025

How to analyse single-cell proteomics data and focus on the underlying biology? Your results are only as good as your method and software.

Mass spectrometry-based single-cell proteomics (SCP) has become a credible player in the single-cell omics arena thanks to substantial technical improvements that have pushed the boundaries of sensitivity and throughput. But what should one do once the precious data have been acquired, often at great cost? Reviewing the SCP literature doesn't provide much help, as every lab tends to run their own in-house, either overly complex or unrealistically trivial and undocumented analysis pipeline. When facing complex data, best is to start with simpler but principled analyses approaches, such as the scplainer method. The goal of scplainer is to move the tension point from how to process SCP data to explain it in the light of the biological question. In this talk, I will use SCP to illustrate how to approach, as a bioinformatician, complex data and it's underlying biological complexity, emphasising the role of research software engineering and computational science.

Slides: https://lgatto.github.io/pub/2025CompSysBio.pdf

Your results are only as good as your method and software.

My analysis is only as good as the explanation and the software to go with it.

Prof Susan Holmes

Is a computational researcher coding doing research?

Is a computational researcher coding doing research?

Better Software, Better Research

Software Sustainability Institute

Is a computational researcher coding doing research?

Better Software, Better Research

Software Sustainability Institute

What is good software? What is good data analysis?

Is a computational researcher coding doing research?

Better Software, Better Research

Software Sustainability Institute

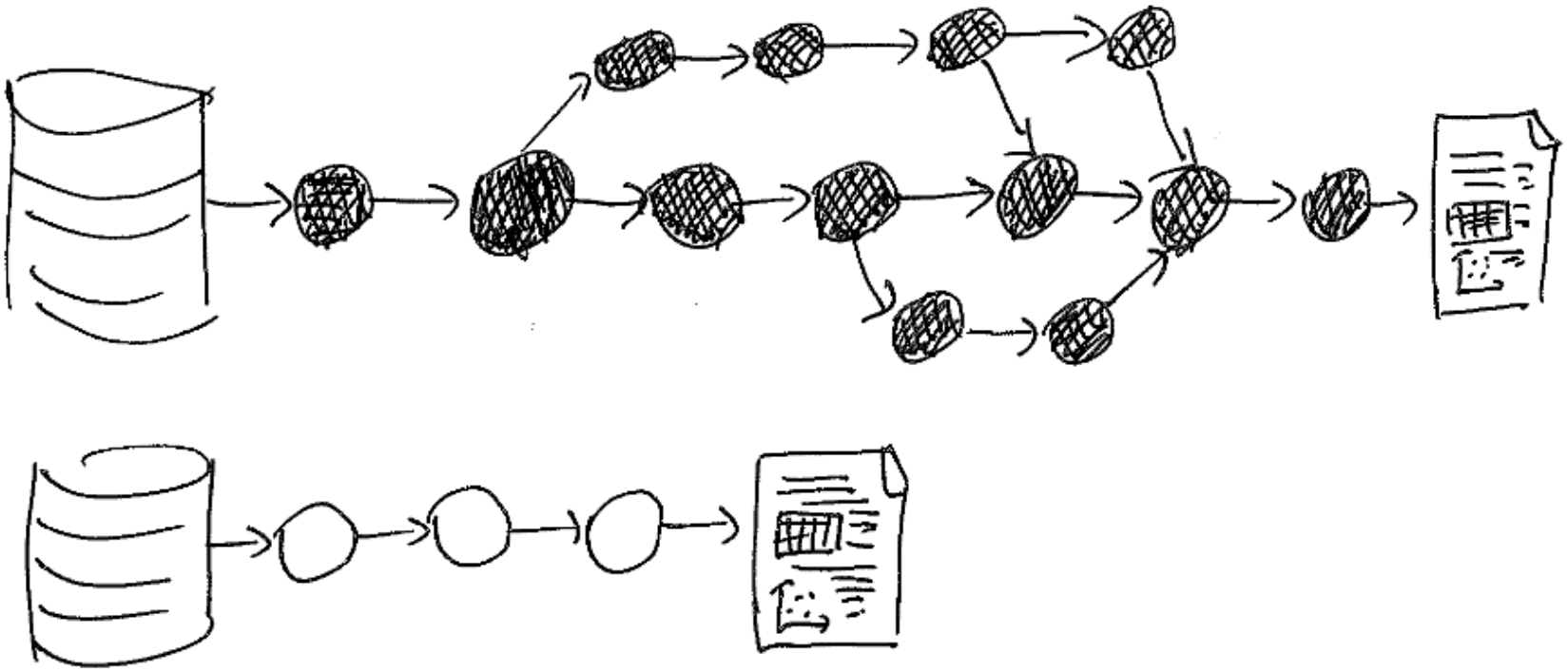
What is good software? What is good data analysis?

Your results are only as good as your method, software and users.

Outline

What is a good data analysis?

Simpler is better



- Data analysis should be as simple as possible, but no simpler.
- Data analysis should be as complex as needed, but not more complex.

Software for data analysis

Software for data analysis

- Compose simple pipelines when possible
- Compose more complex pipelines when necessary
- Enable transparency and reproducibility

Users!

Users!

1. knowledgeable in MS-based (single-cell) proteomics
2. has basic knowledge of data analysis
3. some R (or Python, ...) experience

Outline

Single-cell technologies unravel cellular heterogeneity

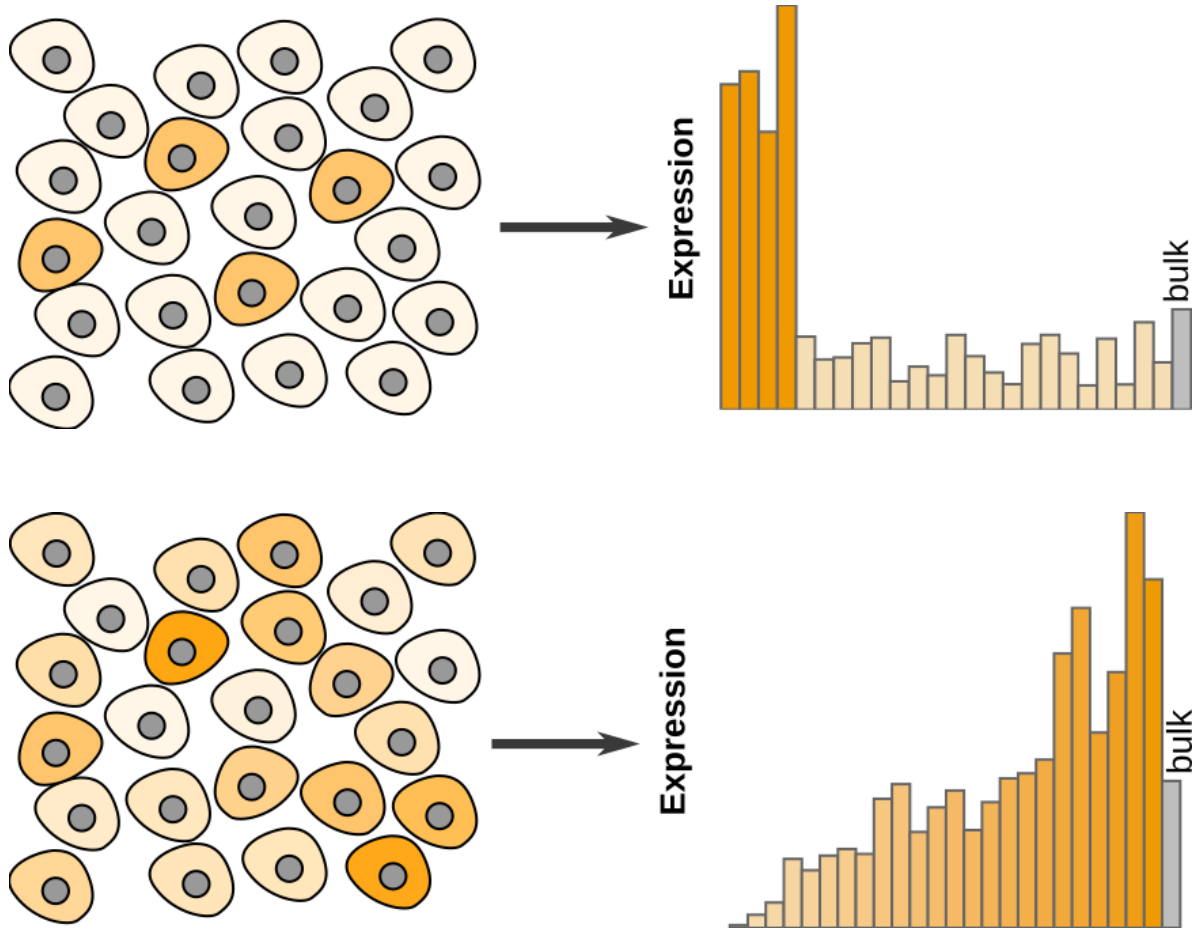


Figure: Cell types and cell states, subpopulation identification, differentiation trajectories (in the absence of known markers).

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**

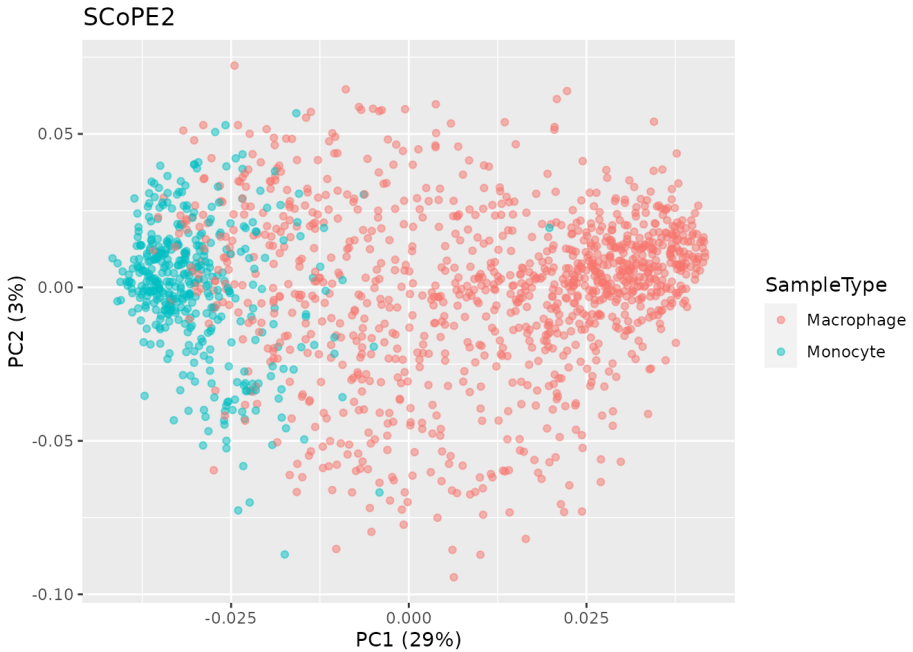
- RNA $\rightarrow$ intention vs. Protein $\rightarrow$ action
- Inference of direct regulatory interactions with minimal assumptions (Slavov, 2022; Hu et al., 2023).
- Post-translational modifications

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**



Methods

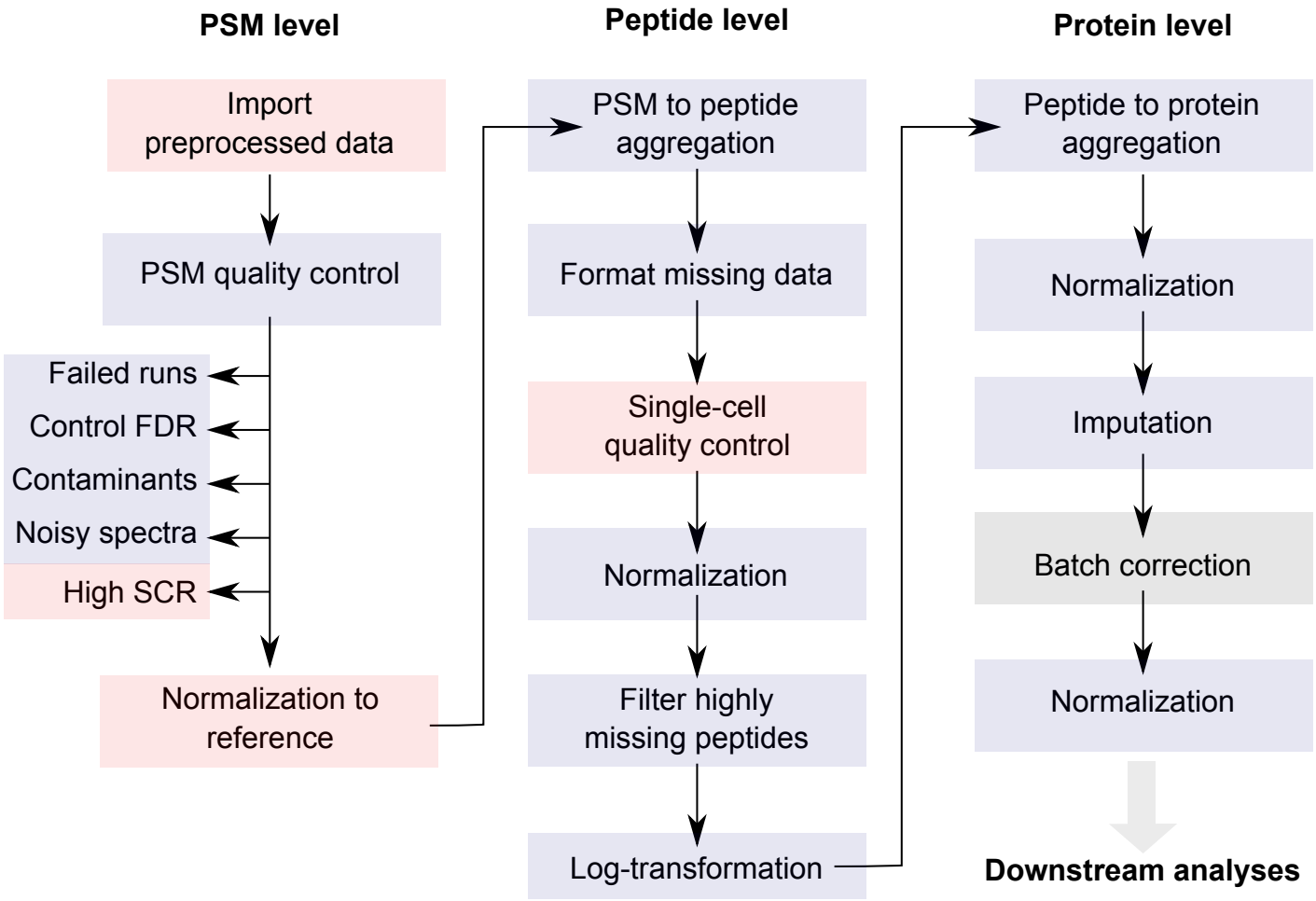


Figure: **Overview of the key steps performed in the SCoPE2 pipeline** (Vanderaa and Gatto, 2021). Blue boxes: **QFeatures**. Red boxes: **scp**. Gray box: **sva::ComBat**.

Outline

Challenge 1: batch effects

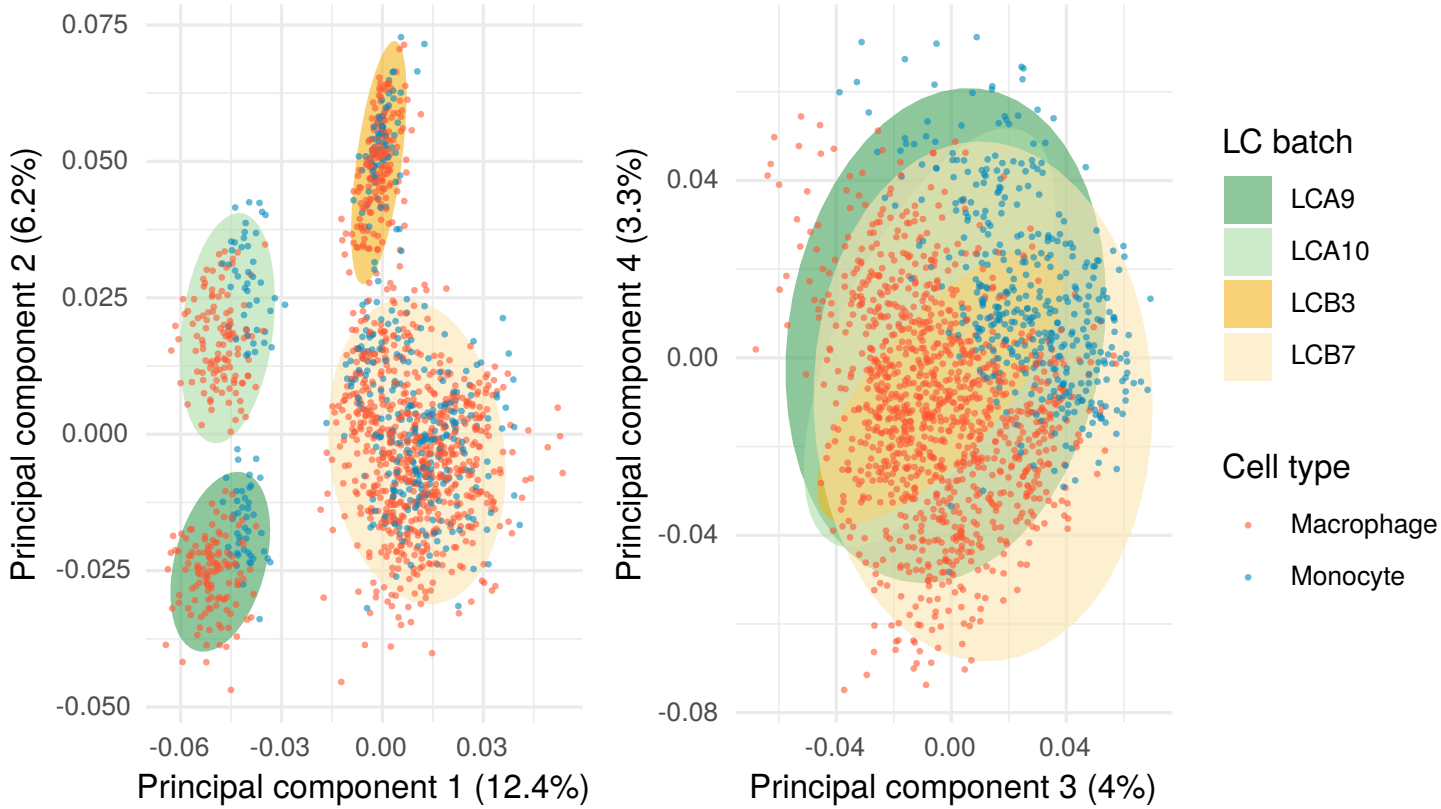


Figure: PCA for the first four components. Each point represents a single-cell and is colored according to the corresponding cell type (Vanderaa and Gatto, 2021).

Challenge 2: missing data

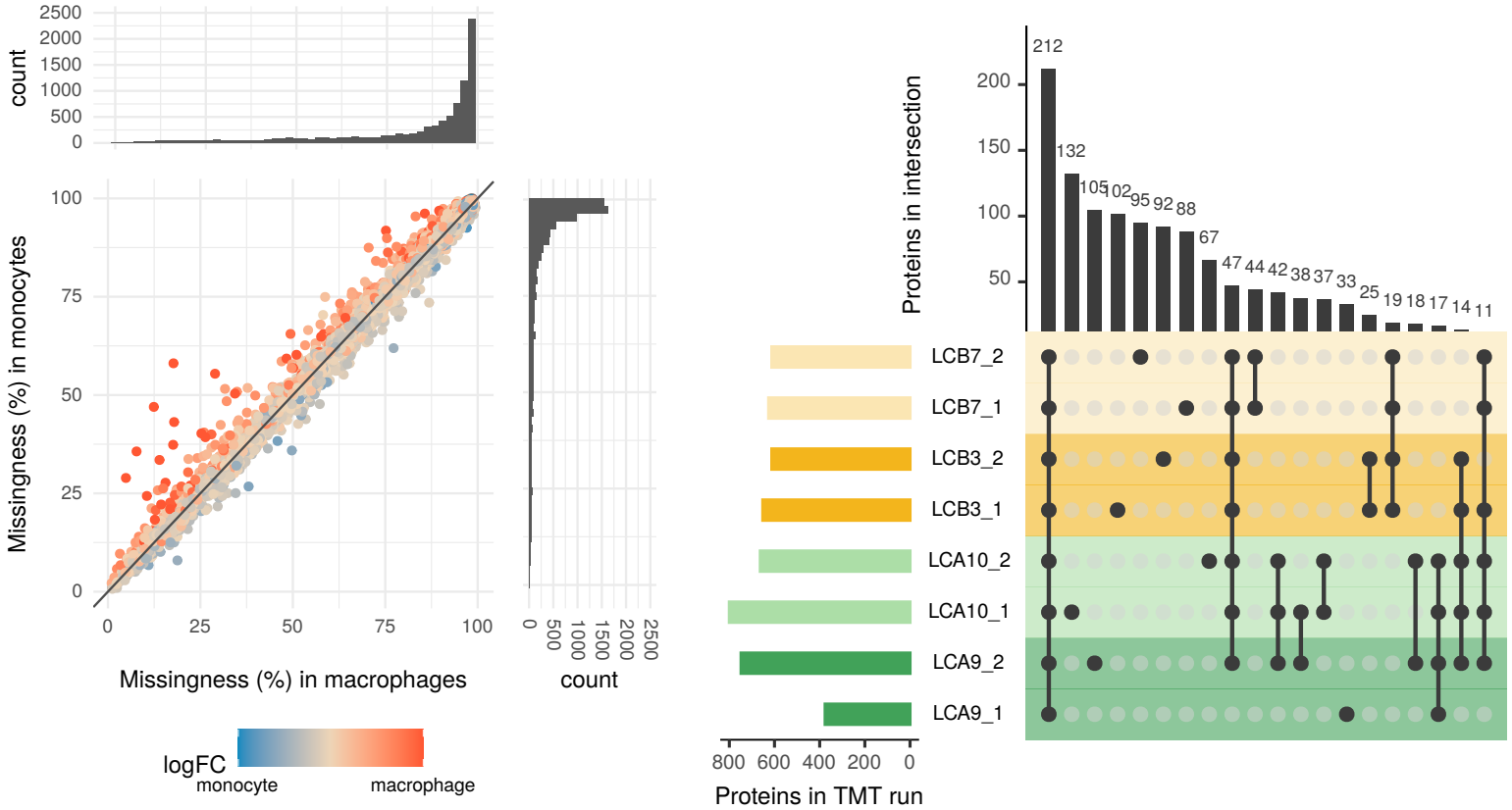


Figure: Missing data is the consequence of biological and technical components (Vanderaa and Gatto, 2021, 2023b).

Challenge 3: 1 + 2

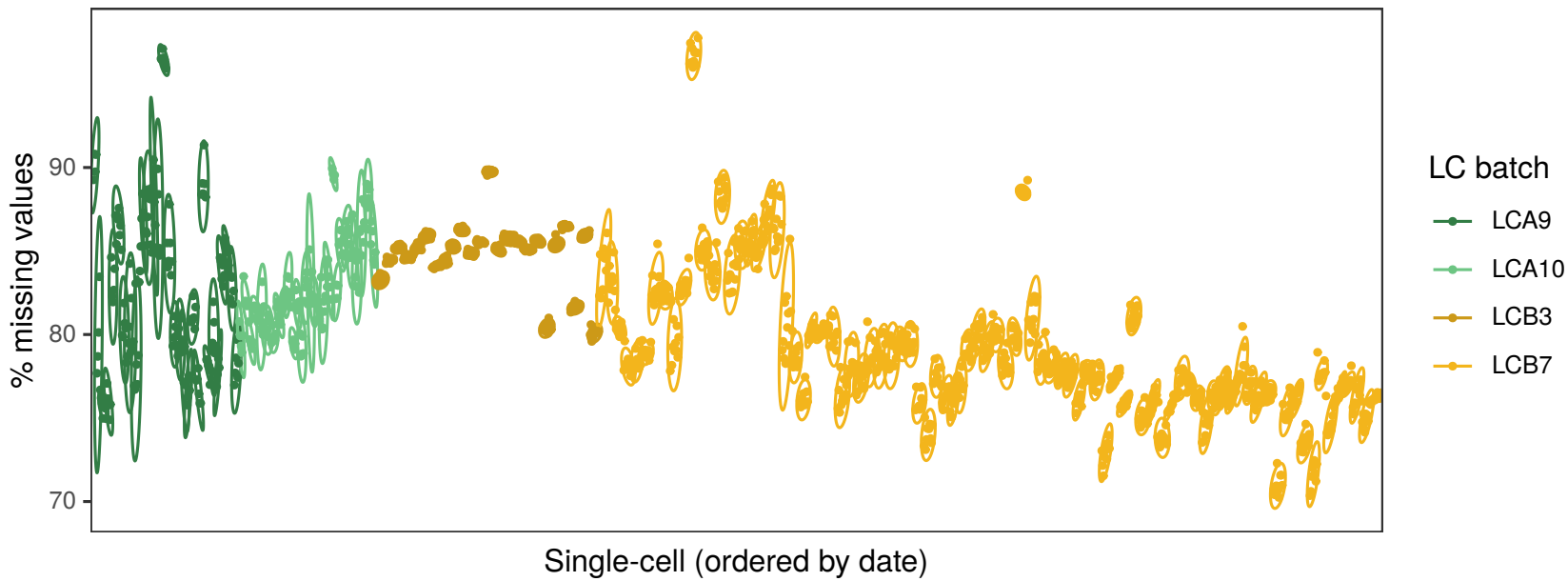


Figure: Influence of batch on data missingness (Vanderaa and Gatto, 2021).

Data analyses review

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

Figure: **SCP.replication**: systematic reproduction/replication of published SCP studies using the **scp** package - **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

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

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

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

(2) Quantify the effects' contributions

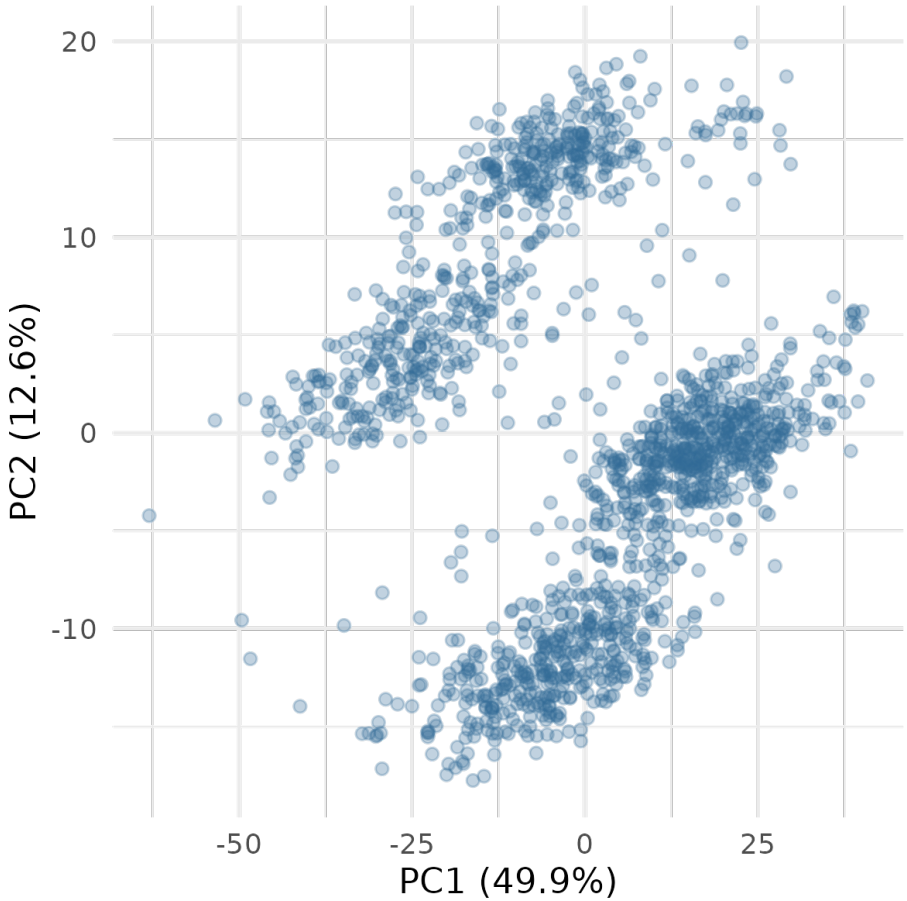
(3) Principal Component Analysis

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**



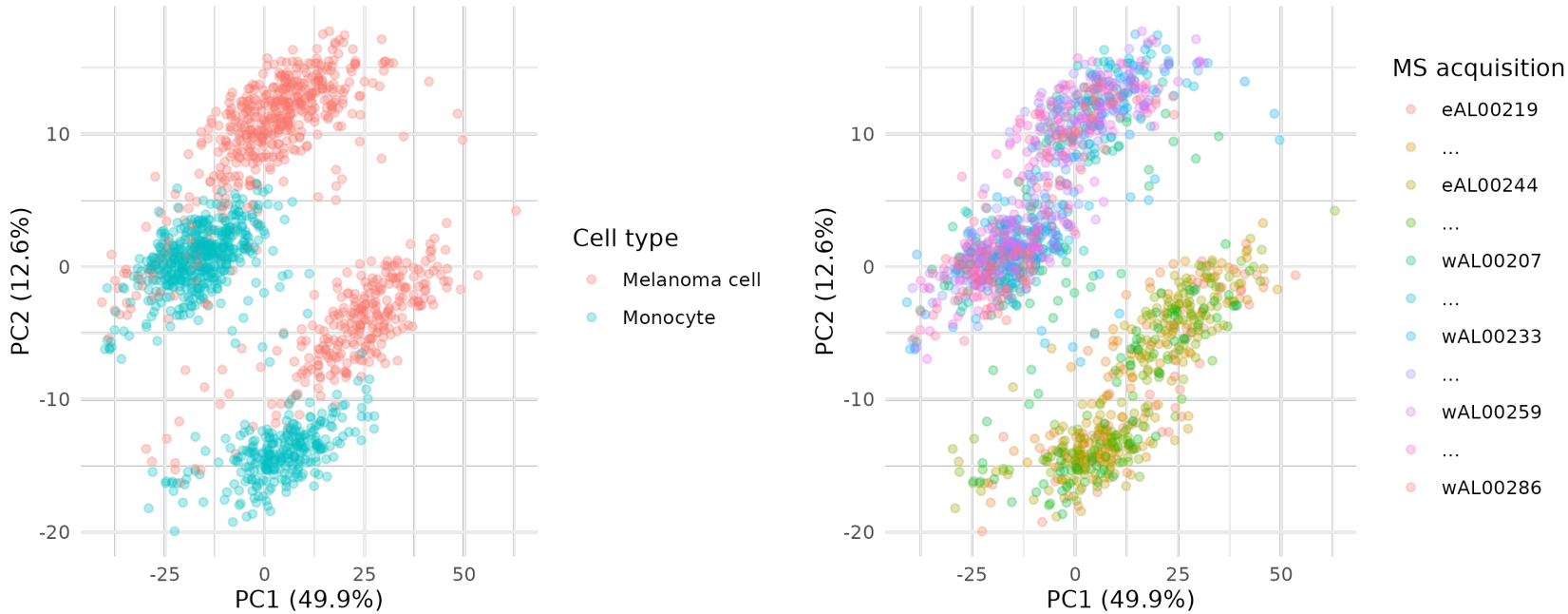


Figure: Melanoma cells and monocytes (left) acquired across multiple acquisition batches (right) (Leduc et al., 2022).

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

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

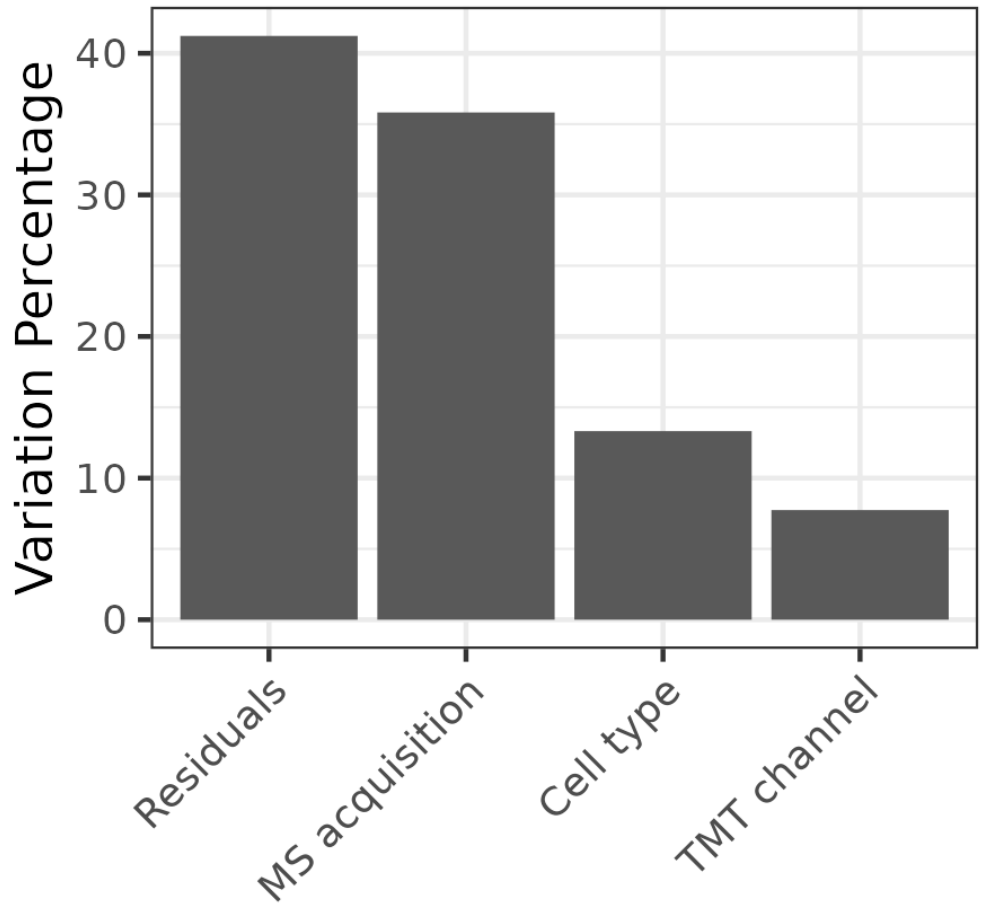
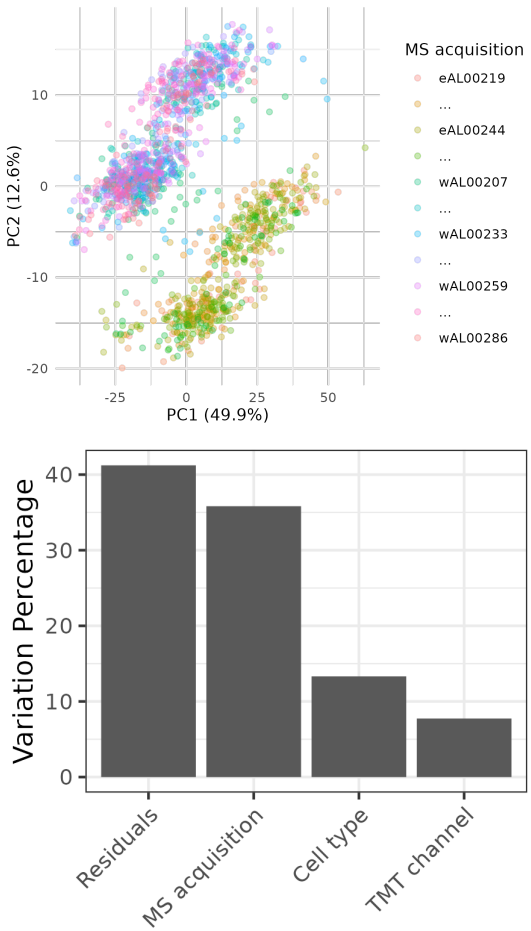


Figure: We are now in a position to **quantify known and unknown effects**: percentages of explained variances of our explained (known) and unexplained (residuals) effects. NB: low biological variance $\neq$ low quality!

PCA on effect matrices

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



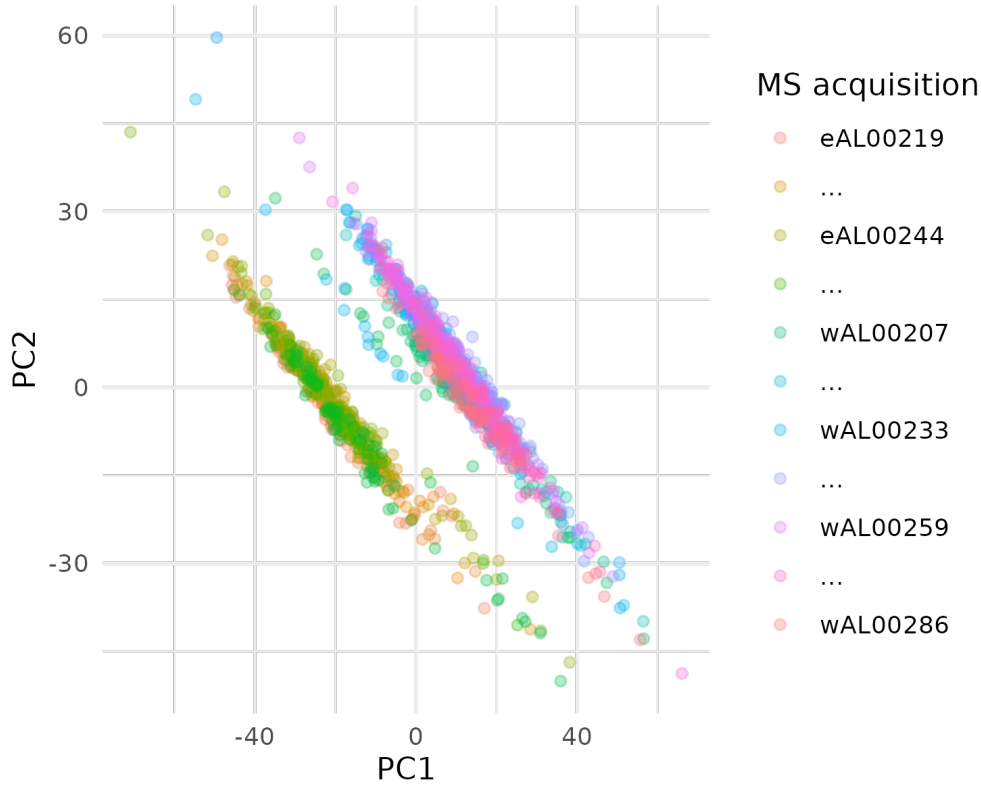
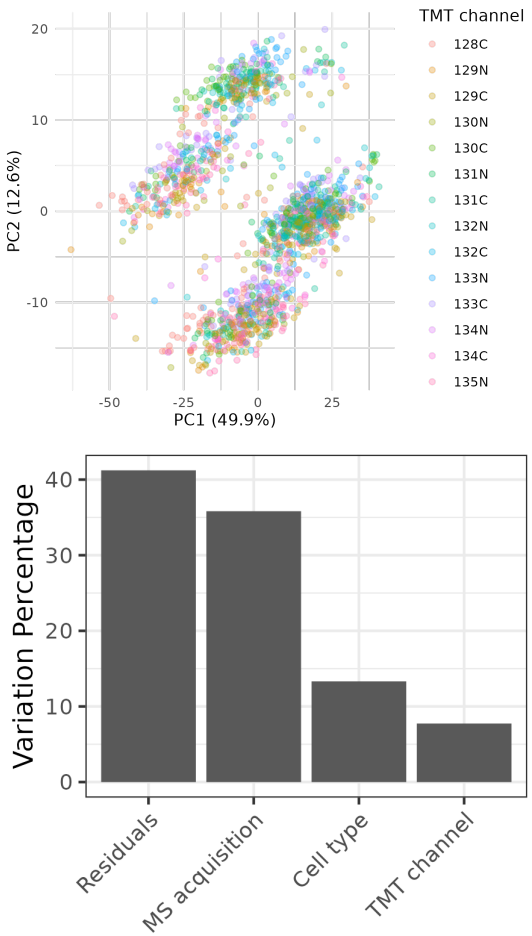


Figure: PCA on the **MS acquisition** effect matrix.

PCA on effect matrices

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



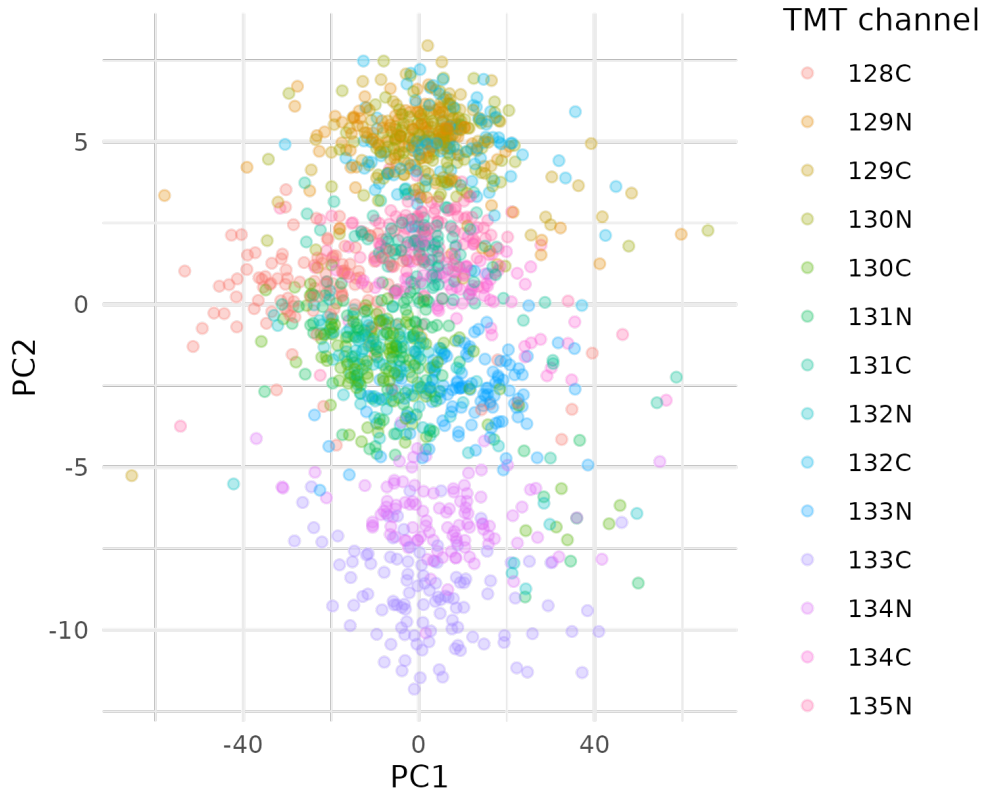
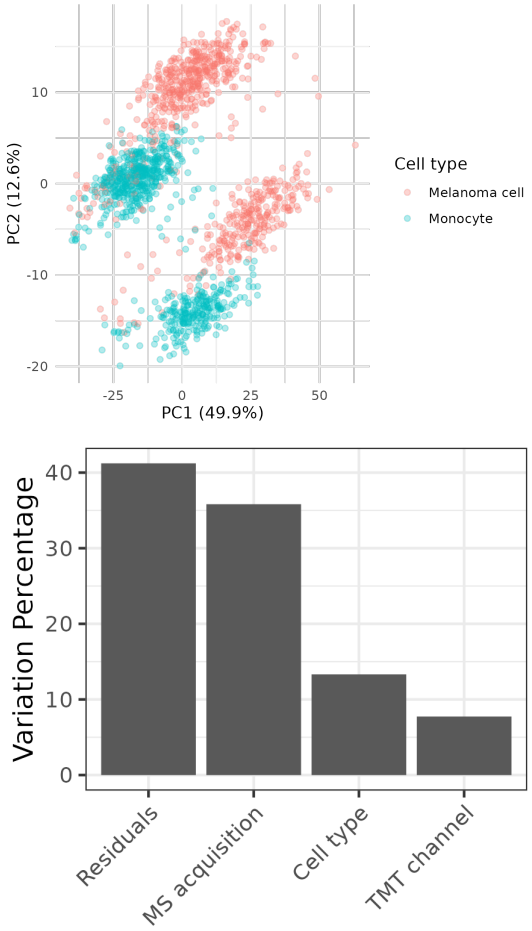


Figure: PCA on the **TMT channel** effect matrix.

PCA on effect matrices

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



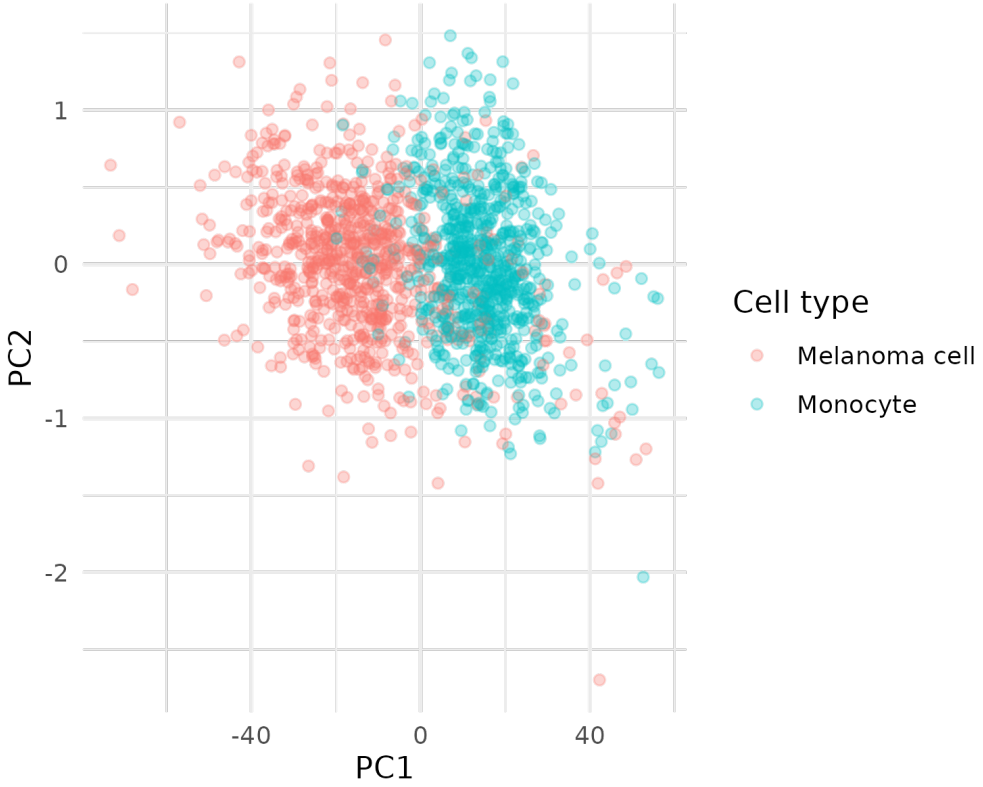
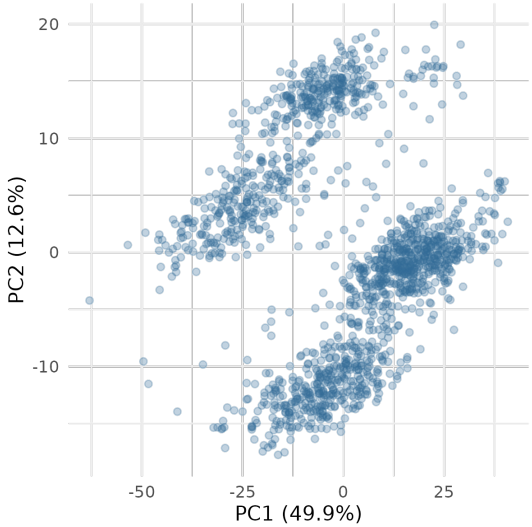
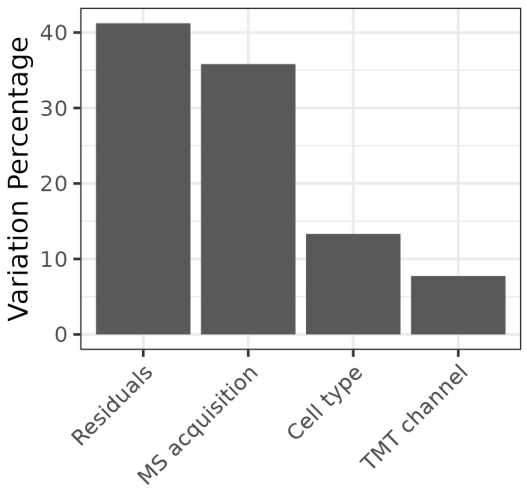


Figure: PCA on the **Cell type** effect matrix.

PCA on effect matrices

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





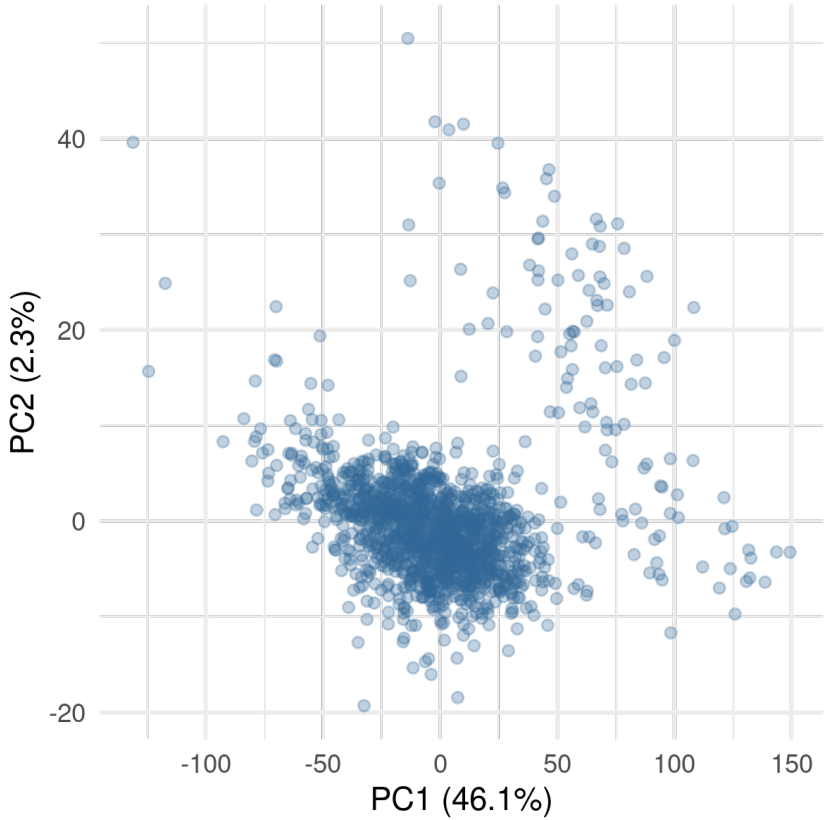
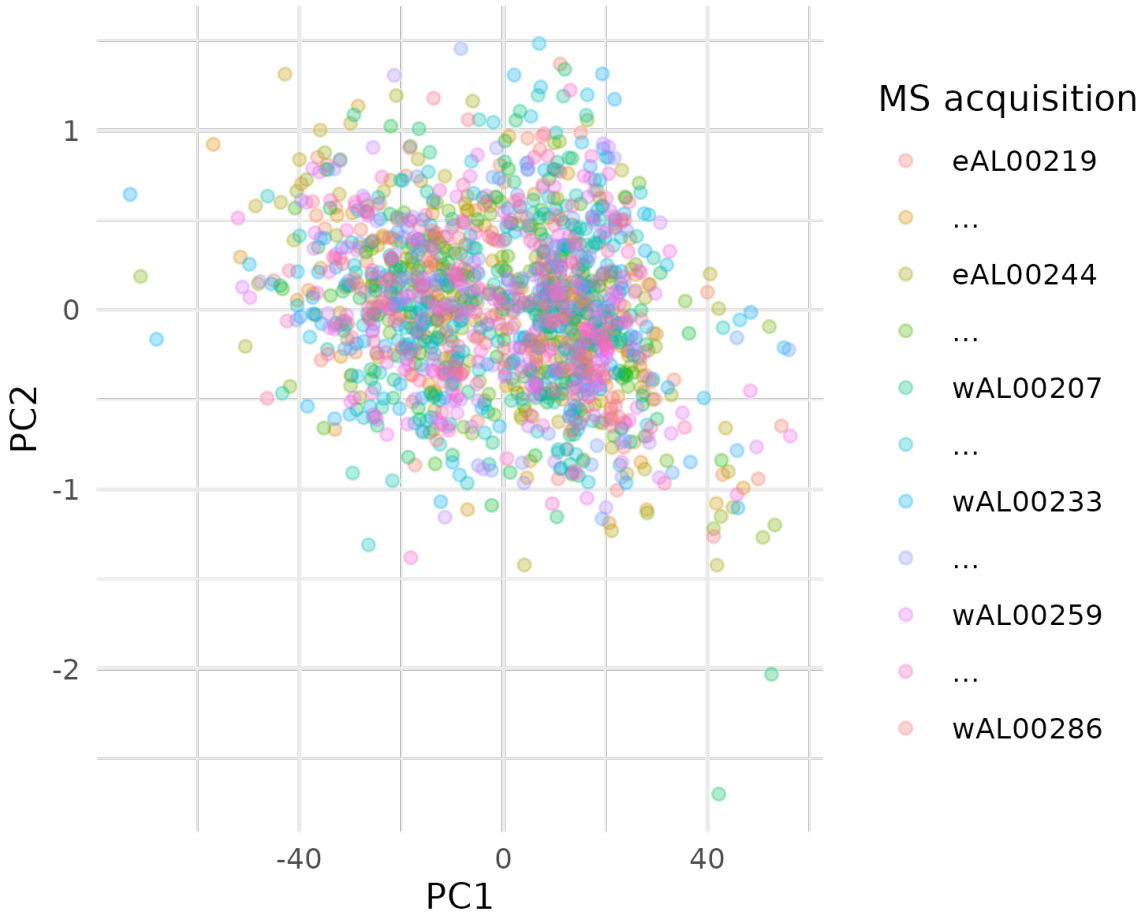


Figure: PCA on the **residuals** effect matrix.

Does it work: negative control

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

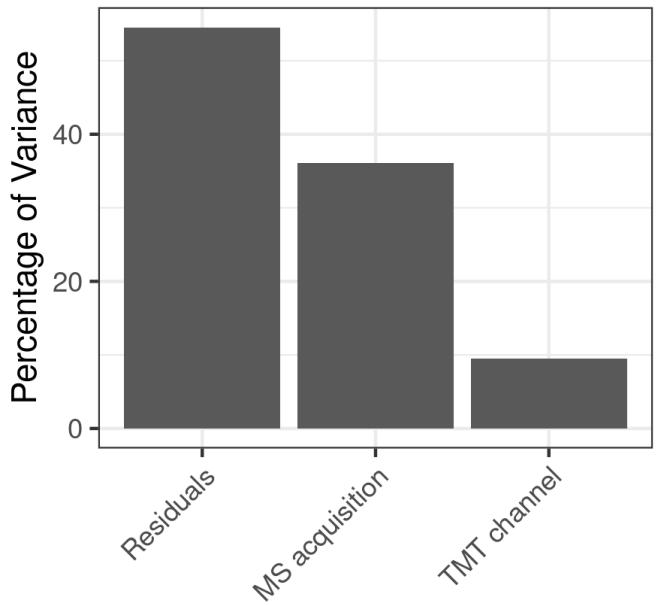


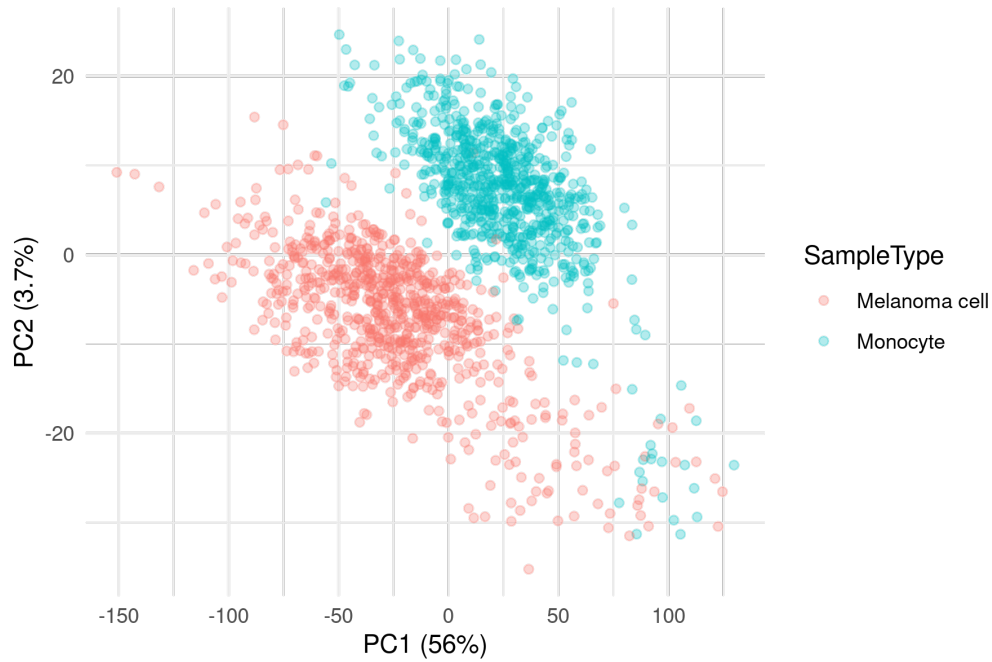
Does it work: positive control

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

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

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

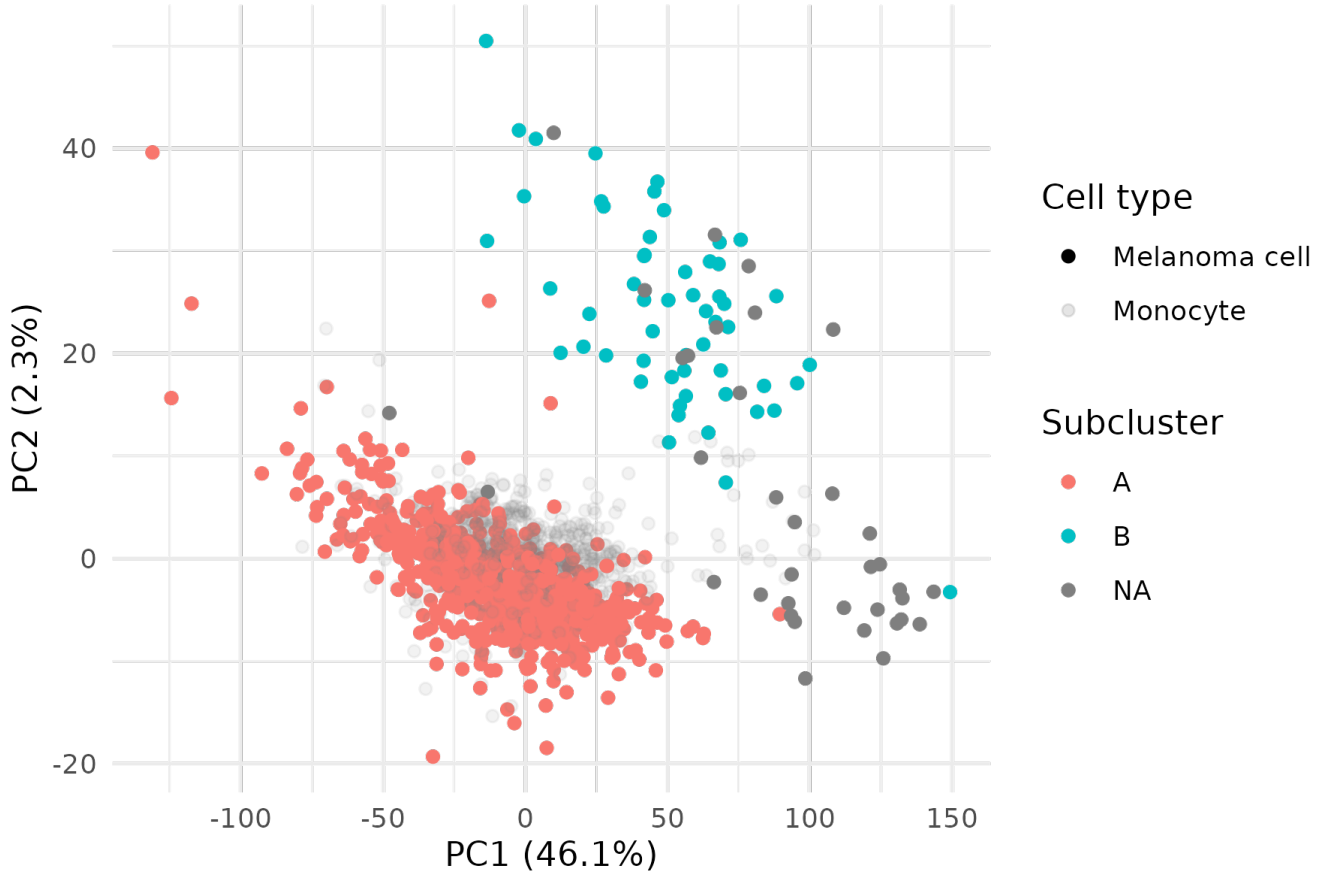


Figure: **Melanoma subpopulations**: transcriptomic signature associated with a cell state that is more likely to resist treatment by the cancer drug vemurafenib (clusters A and B from Leduc et al. (2022)).

Outline

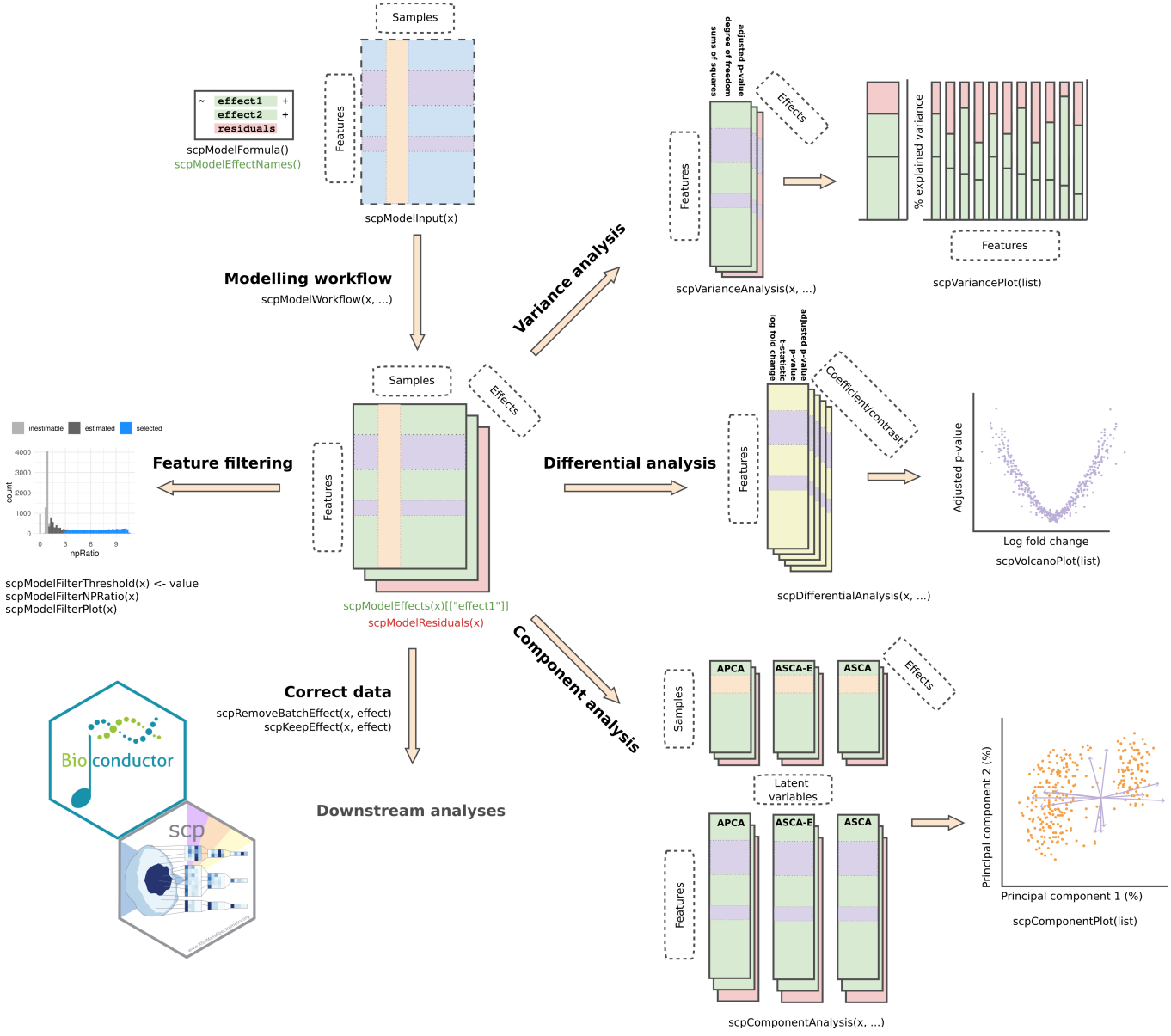


Figure: scp package - scplainer: using linear models to understand mass spectrometry-based single-cell proteomics data (Vanderaa and Gatto, 2025).

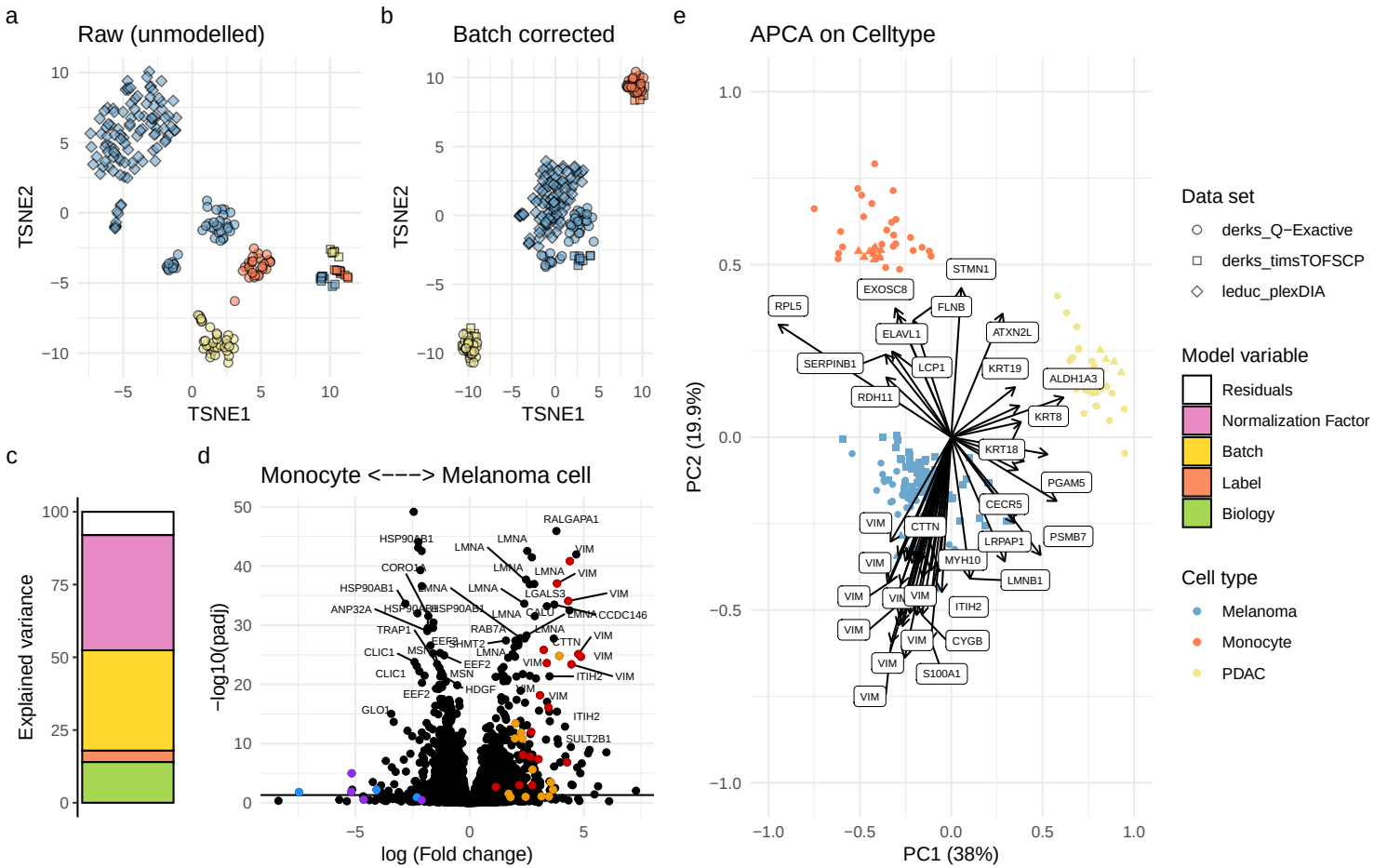


Figure: scplainer – variance, differential and component analysis, integration

What are best RSE practice?

- ...
- ...
- ...

Our software

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

What are best RSE practice?

- Coding practice, style guide, design principles, community/ISO/IEC standards, unit and integration testing, CI, code/peer review, automation, version control, ...
- Documentation, tutorials, courses, user support, ...
- Supportive community, code of conduct, ...

Our software

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

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.
- **Limitation**: multi-patient/condition designs - mixed effects (Sticker et al., 2020) and pseudo-bulking.

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.
- **Limitation**: multi-patient/condition designs - mixed effects (Sticker et al., 2020) and pseudo-bulking.

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

References I

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.

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.

References II

Adriaan Sticker, Ludger Goeminne, Lennart Martens, and Lieven Clement. Robust summarization and inference in proteome-wide label-free quantification. *Mol. Cell. Proteomics*, 19(7):1209–1219, July 2020.

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. *Genome Biol.*, 26(1):237, August 2025.

UCLouvain

Acknowledgments

- **Computational Biology and Bioinformatics** lab, de Duve Institute, UCLouvain – lgatto.github.io/cbio-lab.
- Dr Christophe Vanderaa (CBIO, now UGent).
- Prof Nikolai Slavov, Parallel Squared Technology Institute.

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

Discussion points

- Your results are only as good as your method, software and users.
- Is a computational researcher coding doing research?
- What is good software? What is good data analysis?
- Should all software meet the highest standard? Should every piece of research be 100% reproducible?
- What about LLM-generated code?

Slides: https://lgatto.github.io/pub/2025CompSysBio.pdf