scpdata: a data package for single-cell proteomics

Christophe Vanderaa, Laurent Gatto

Computational biology and bioinformatics, de Duve Institute, UCLouvain

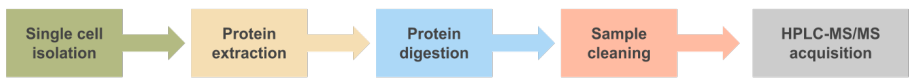
christophe.vanderaa@uclouvain.be

Summary

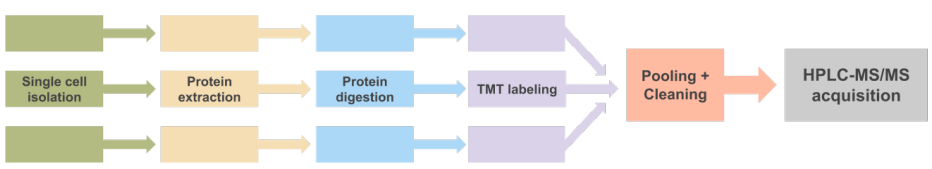
Recent advances in sample preparation, processing and mass spectrometry (MS) have allowed the emergence of MS-based **single-cell proteomics** (SCP). However, bioinformatics tools to process and analyze these new types of data are still missing. In order to boost the development and the benchmarking of SCP methodologies, we are developing the `scpdata` experiment package. The package will distribute published and **curated** SCP data sets in **standardized Bioconductor** format.

Introduction

There are two main pipelines able to generate MS-SCP data: **nanoPOTS pipeline** (Zhu et al., 2018, [1]) runs label-free proteomics for single cells. The **throughput is low** (± 10 samples/day), but it achieves **accurate peptide quantification**.



SCoPE pipeline (Budnik et al., 2018, [2]) adapts TMT-based proteomics to single-cells. The **throughput is higher** (± 5 samples/hour), but it suffers from **presence of chemical noise**.



Data manipulation

The Bioconductor class `MSnSet` is a reliable framework for **standard and systematic** quantitative data processing. Below, we have reproduced the analysis pipeline from [3]:

```
data("specht2019_peptide")
specht2019_peptide %>%
  scp_normalize_stat(what = "row", mean, "-") %>%
  scp_aggregateByProtein() %>%
  scp_normalize_stat(what = "column", median, "-") %>%
  scp_normalize_stat(what = "row", mean, "-") %>%
  imputeKNN(k = 3) %>%
  batchCorrect(batch = "raw.file", target = "celltype") -> scpd
```

Data quality control

When developing the SCoPE technology, the Slavov lab also suggested some quality control (QC) measures and visualizations [4] (Figure 1). The `scpdata` package provides the framework to generalize those metrics.

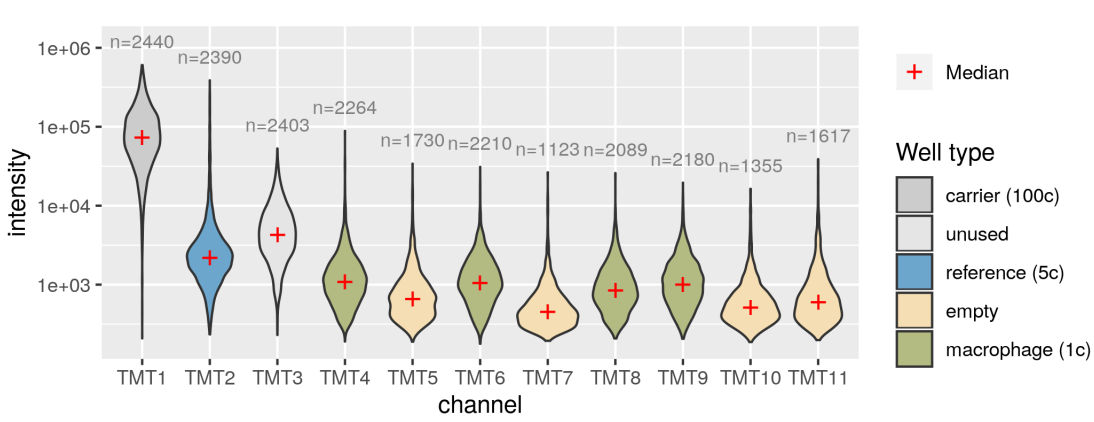


Figure 1: **MS intensity distributions per channel at peptide level.** Contamination peptides or peptides with a low identification score were removed. Data taken from run `190222S_LCA9_X_FP94BF` published in [3]. n: number of non-missing peptides.

Content of the package

`scpdata` contains SCP data sets formatted as `MSnbase::MSnSet` objects [5]. The package provides data at **peptide** and **protein** level. **Help files** are provided for every data set. Available data sets are listed using `scpdata()`.

Item	Title
dou2019_1_protein	FACS + nanoPOTS + TMT multiplexing: HeLa digests (Dou et al. 2019)
dou2019_2_protein	FACS + nanoPOTS + TMT multiplexing: testing boosting ratios (Dou et al...
dou2019_3_protein	FACS + nanoPOTS + TMT multiplexing: profiling of murine cell populations ...
specht2018_peptide	SCoPE-MS + mPOP lysis upgrade: Master Mix 20180824 (Specht et al. 2018)
specht2019_peptide	FACS + SCoPE2: comparing macrophages against monocytes (Specht et al. 2019
specht2019_peptide2	FACS + SCoPE2: comparing macrophages against monocytes (Specht et al. 2019
specht2019_protein	FACS + SCoPE2: comparing macrophages against monocytes (Specht et al. 2019

Conclusion

MS-based SCP is still in its infancy. Nevertheless, the `scpdata` experiment package offers a growing repository of curated data ideally suited for **method benchmarking** and **data QC**. This will enable us to develop new methodologies to tackle the current hurdles that MS-SCP faces: missing data, batch effect, and high dimensionality.

Benchmarking

`scpdata` also offers an ideal environment for benchmarking. It will contain a wide variety of MS-SCP data sets from **well-defined synthetic standards** to **real biological samples**. Different methods can be compared using **objective benchmarking metrics** or **visualization** with dimension reduction (Figure 2).

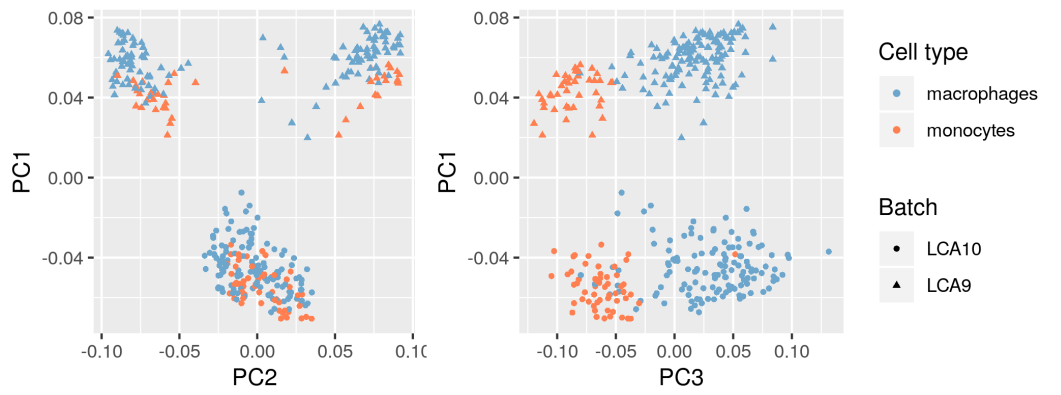


Figure 2: **PCA plot of peptide expression data.** Macrophages and monocytes are well separated in the third principal component. However, the first and second components are driven by batch effects. LCA10 and LCA9 are two chromatographic batches. The PCA was performed using the NIPALS algorithm.

Problems to tackle

Batch effects

Batch effects are inherent to MS-SCP data since many samples/cells have to be distributed across **different MS runs**. This leads to major biases in the data (Figure 2).

Missingness

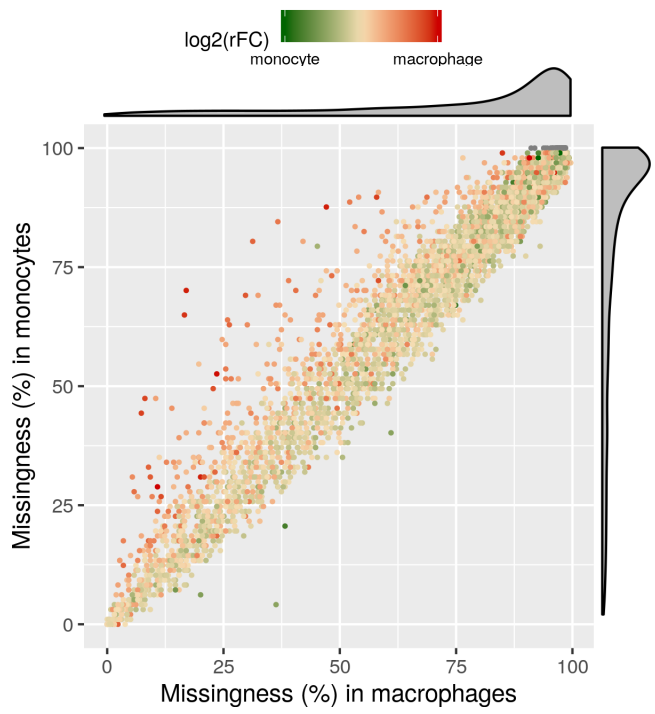


Figure 3: **Distribution of missing data in monocytes against macrophages.** The average missingness is ± 75 %. Color indicates the log2 fold change of **macrophages** over **monocytes** relative expression. Data from [3].

Curse of dimensionality

Although current acquisition pipelines produce data sets of **thousands of peptides x hundreds of cells**, it is expected that new technological advances might raise the dimensionality 100 fold [3]. This is a challenge for the **statistical analyses** and for the **software optimization**. Possible solutions should be inspired from current achievements in single cell transcriptomics.

This work is funded by an Aspirant FRS-FNRS fellowship awarded to Christophe Vanderaa. The poster is available at https://github.com/cvanderaa/EuroBioc2019-Poster.

References

[1] Y. Zhu, P. D. Piehowski, R. Zhao, J. Chen, Y. Shen, R. J. Moore, A. K. Shukla, V. A. Petyuk, M. Campbell-Thompson, C. E. Mathews, R. D. Smith, W.-J. Qian, and R. T. Kelly, "Nanodroplet processing platform for deep and quantitative proteome profiling of 10-100 mammalian cells," *Nat. Commun.*, vol. 9, p. 882, Feb. 2018.

[2] B. Budnik, E. Levy, G. Harmange, and N. Slavov, "SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation," *Genome Biol.*, vol. 19, p. 161, Oct. 2018.

[3] H. Specht, E. Emmott, T. Koller, and N. Slavov, "High-throughput single-cell proteomics quantifies the emergence of macrophage heterogeneity." June 2019.

[4] G. Huffman, H. Specht, A. T. Chen, and N. Slavov, "DO-MS: Data-Driven optimization of mass spectrometry methods." Jan. 2019.

[5] L. Gatto and K. S. Lilley, "MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation," *Bioinformatics*, vol. 28, pp. 288–289, Jan. 2012.