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]:

```r
1 2 3 4 5
1) data("specht2019_protein")
2) specht2019_protein %>%
3) scp_normalize_stat(what = "row", mean, 
4) scp_aggregateByProtein() %>
5) scp_normalize_stat(what = "column", median, 
```

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.

```r
data("specht2019_peptide")
specht2019_peptide %>
scp_normalize_stat(what = "row", mean, 
scp_aggregateByProtein() %>
scp_normalize_stat(what = "column", median, 
scp_aggregateByProtein() %>
```

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()

References