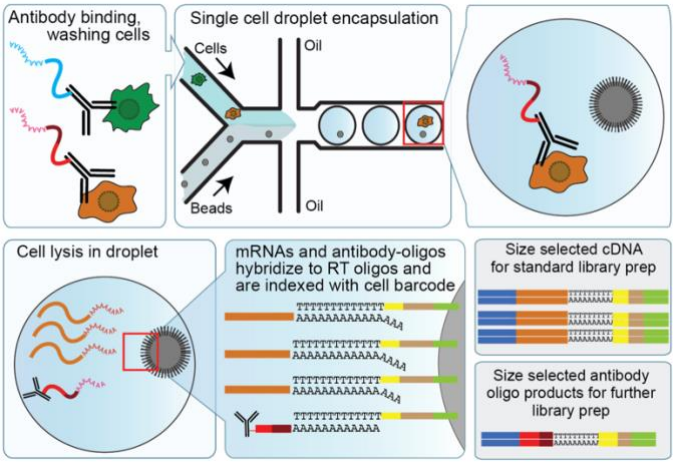
Differential expression analysis for the protein component of CITE-seq data.

CITE-seq is an exciting new technology that allows for the simultaneous quantification of transcripts and extracellular proteins in single-cells. It has already been shown that having the extracellular protein data as an additional layer of information for the identification of cell types that could not be discovered by only using scRNA-seq. In addition, CITE-seq has the promise to increase our understanding of post-transcriptional gene regulation at the single-cell level. In the figure on the right, a schematic representation of a CITE-seq protocol is displayed.



One of the key tasks in the downstream analysis of CITE-seq data is differential expression (DE) analysis. DE analyses have the goal to identify features, in this case transcripts or extracellular proteins, that are differentially abundant between, for instance, different cell types. Currently, very few methods for performing DE analysis on CITE-seq data have been proposed. The most notable method in this context is the Bioconductor R package CiteFuse. For DE analysis, CiteFuse considers the transcript data and protein data separately. For both datatypes, CiteFuse proposes a Wilcoxon rank sum (WRS) test to identify differentially expressed features. However, the WRS test may be suboptimal for DE analyses on these data types. For the transcript data, WRS tests may suffer from the large number of very low and zero counts, which introduces ties in the data. For the protein data, which has much higher counts, it may be more efficient to make some distributional assumptions for the data, i.e. to obtain higher statistical power for identifying DE proteins.

In this assignment, we will mainly focus on DE analysis for the protein component of CITE-seq data. It will mostly be an exploratory analysis, in the sense that we will assess several distributional assumptions for modelling the protein expression levels (i.e. Poisson, negative binomial, …) and explore their goodness-of-fit. In a later stage, this would allow us to propose a model for identifying DE proteins in CITE-seq data that has higher statistical power than the current state-of-the art software.

Preliminary readings:

1. Paper that first introduced the rationale behind CITE-seq experiments: https://doi.org/10.1038/nmeth.4380
2. Paper behind the Bioconductor R package CiteFuse, that implemented, amongst others, a rudimentary algorithm for performing DE analyses on CITE-seq data: https://doi.org/10.1093/bioinformatics/btaa282
   Note; this paper is not solely focused on DE analysis; paragraph 2.6. is the most relevant.

Datasets:

1. The dataset associated with the citeFuse R package and vignette (https://sydneybiox.github.io/CiteFuse/articles/CiteFuse.html#differential-expression-analysis-1)
2. The dataset used in the OSCA workflow (Chapter 20), which can be obtained through:

```
library(DropletTestFiles)
path <- getTestFile("tenx-3.0.0-pbmc_10k_protein_v3/1.0.0/filtered.tar.gz")
```