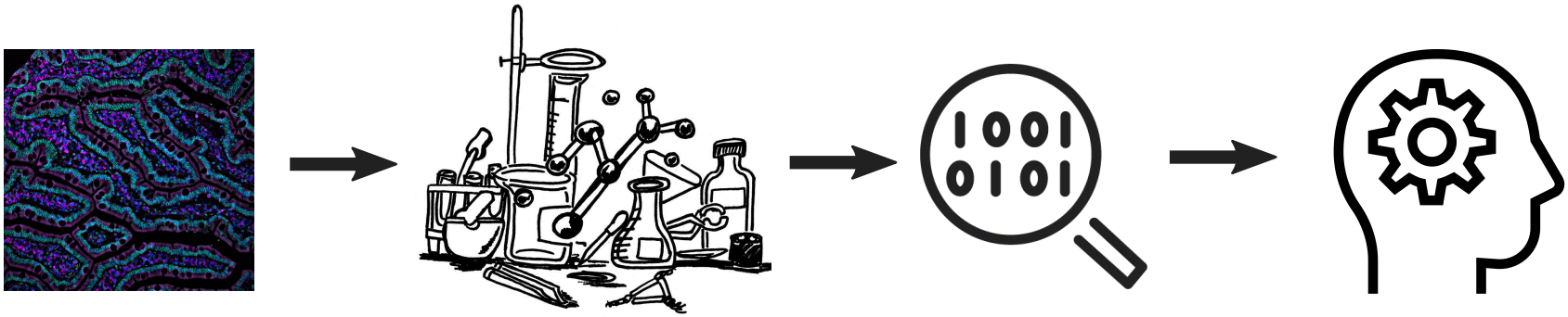
A standardized computational framework for the analysis of mass spectrometry-based single-cell proteomics data

Christophe Vanderaa, Laurent Gatto

December 4, 2020

Bioinformatics

Bioinformatics = understand biology from data



Types of measures:

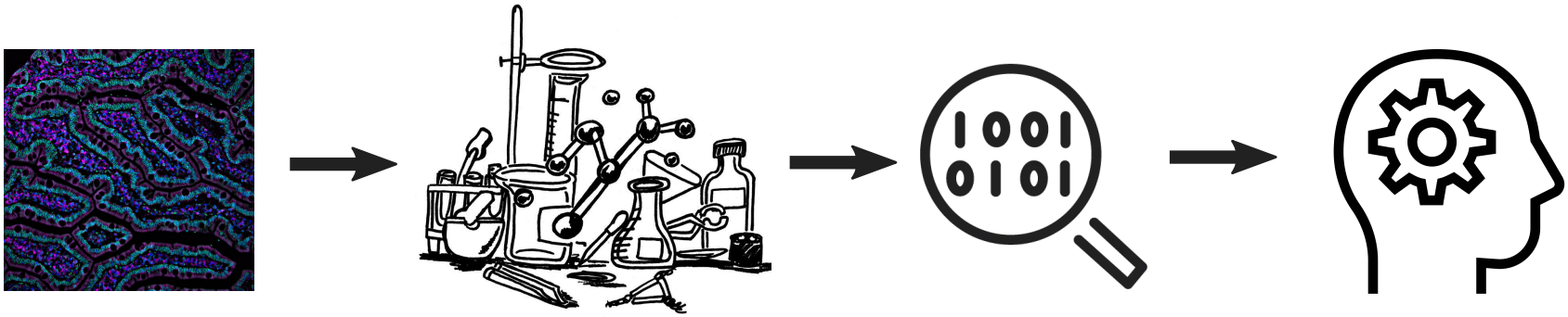
- Sequencing
- Probing
- Microscopy
- Simulation
- ...

Types of levels:

- Epigenomics
- Genomics
- Transcriptomics
- Proteomics
- ...

Bioinformatics

Bioinformatics = understand biology from data



Types of measures:

- **Sequencing**
- Probing
- Microscopy
- Simulation
- ...

Types of levels:

- Epigenomics
- Genomics
- Transcriptomics
- **Proteomics**
- ...

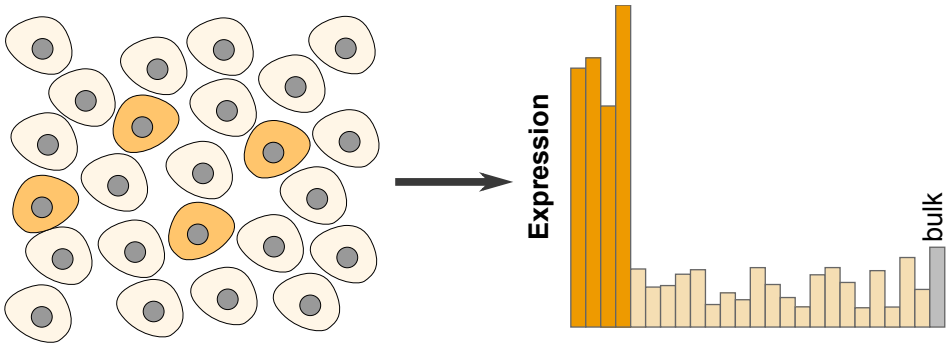
Bulk vs single-cell omics

Bulk omics generates a single observation for highly complex samples

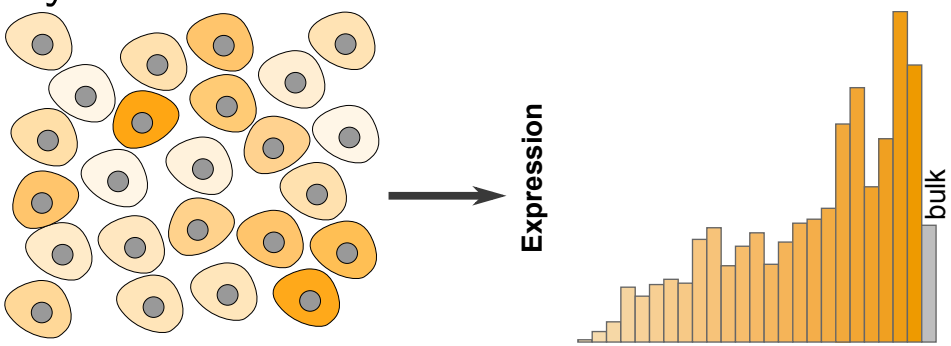
Bulk vs single-cell omics

Bulk omics generates a single observation for highly complex samples

- Subpopulations are missed



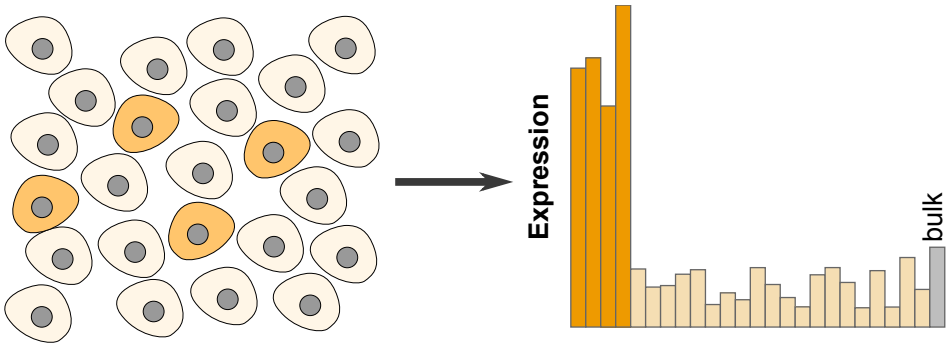
- Dynamic effects are missed



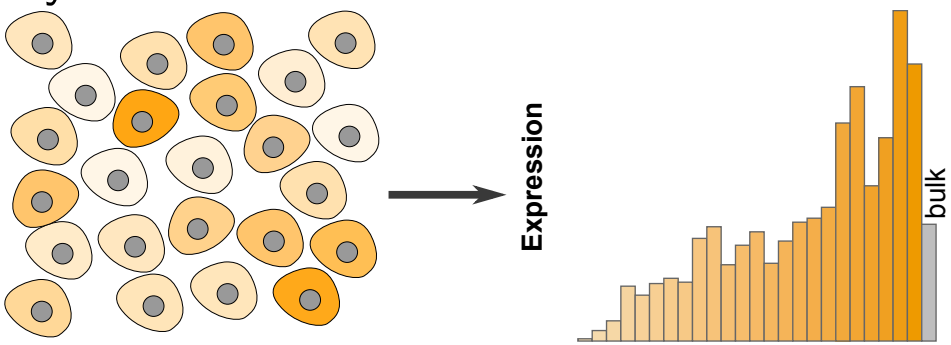
Bulk vs single-cell omics

Bulk omics generates a single observation for highly complex samples

- Subpopulations are missed



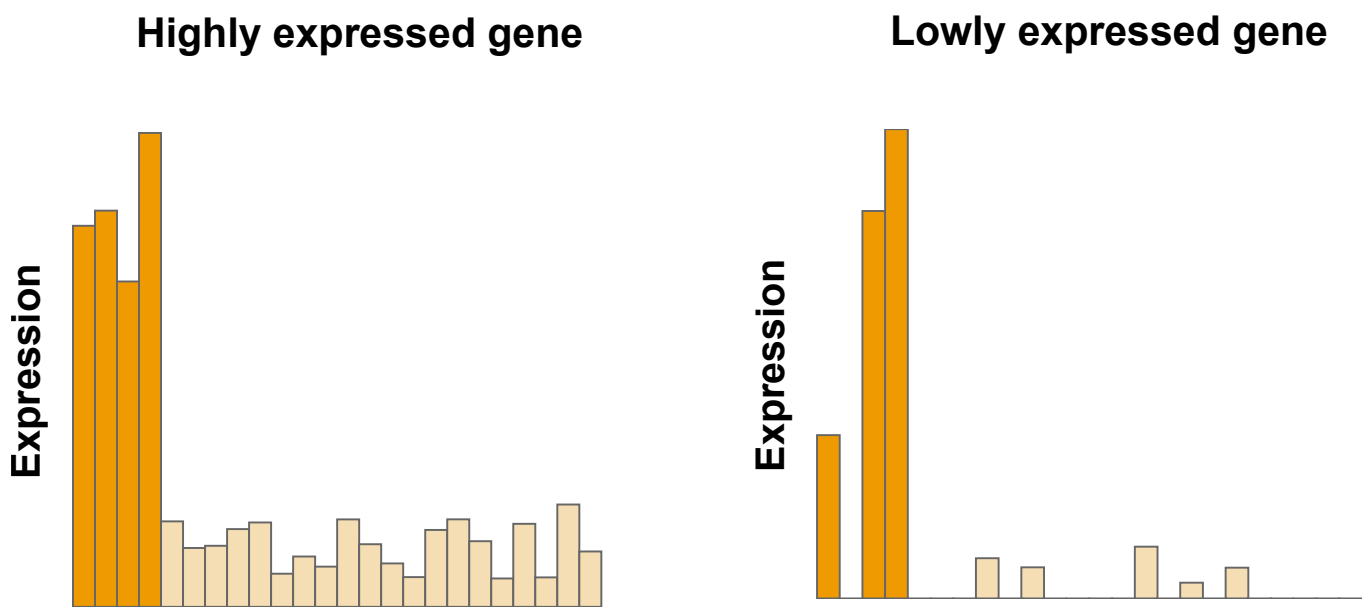
- Dynamic effects are missed



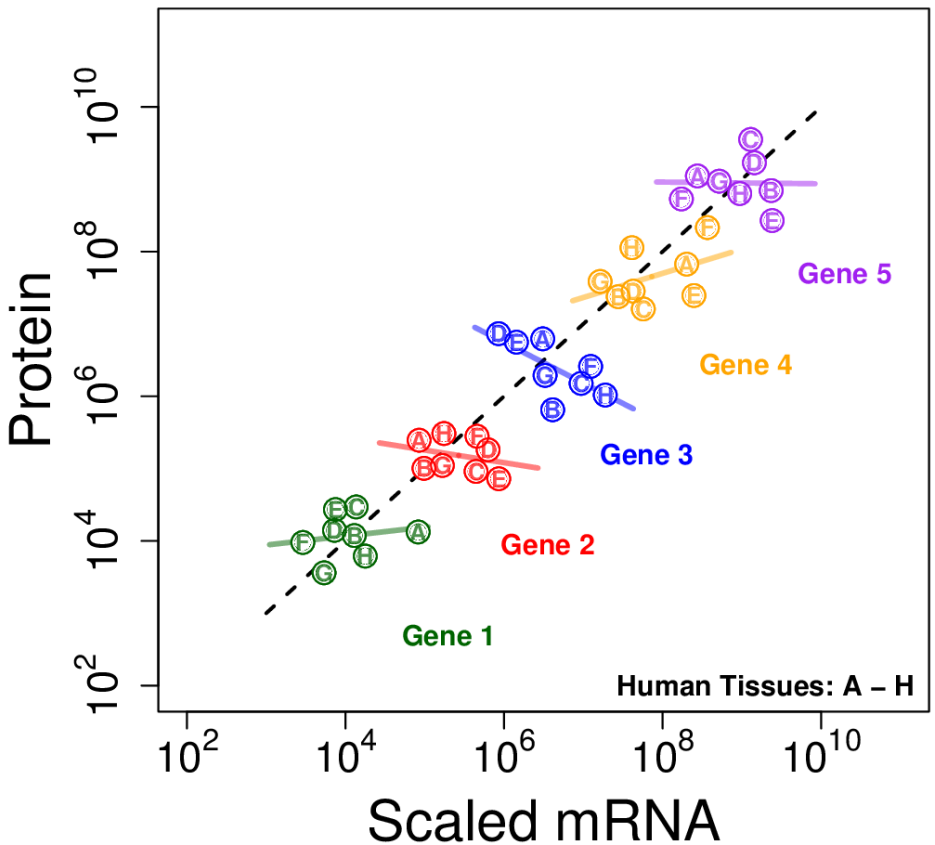
Single-cell omics generate one observation per cell, unlocking new analytical tools

Single-cell challenges

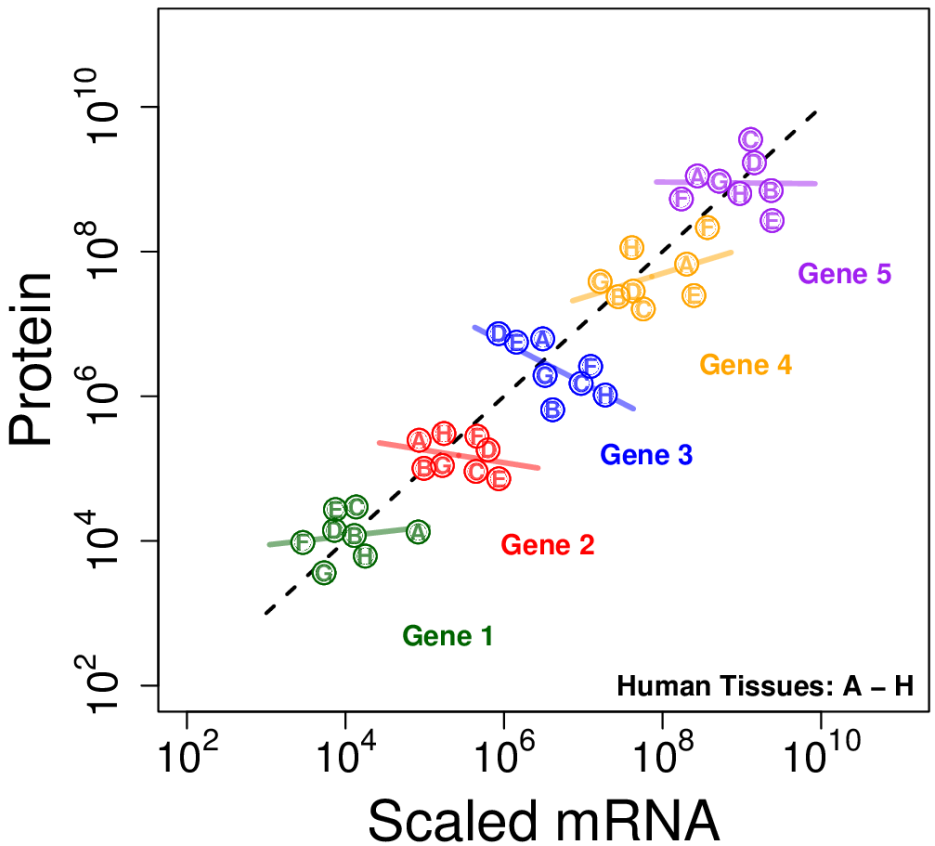
- Technical challenges: automation, minute sample amount, cost per cell
- Computational challenges: big data, dropouts, noise, complex batch effects
- Conceptual challenges: what is a cell type? what is biologically relevant?

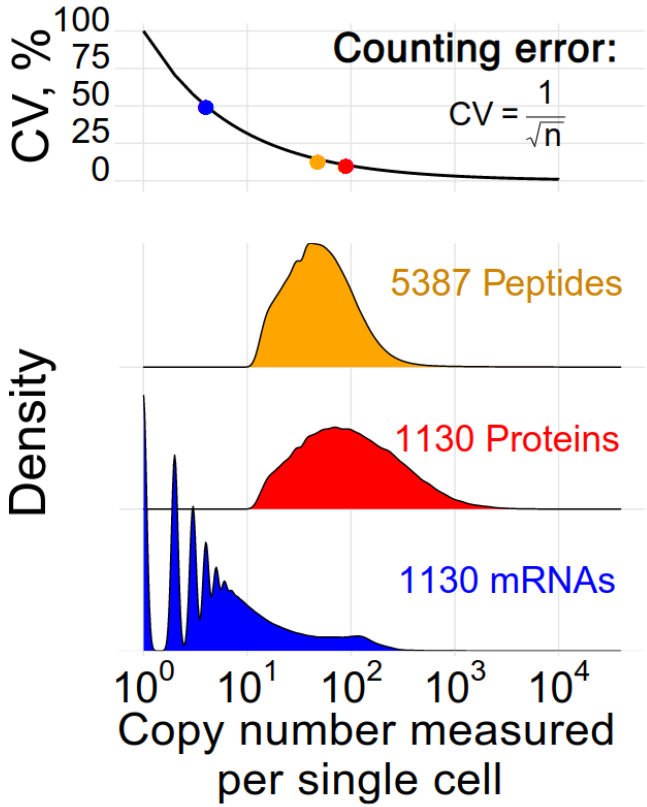


Proteomics vs transcriptomics



Proteomics vs transcriptomics

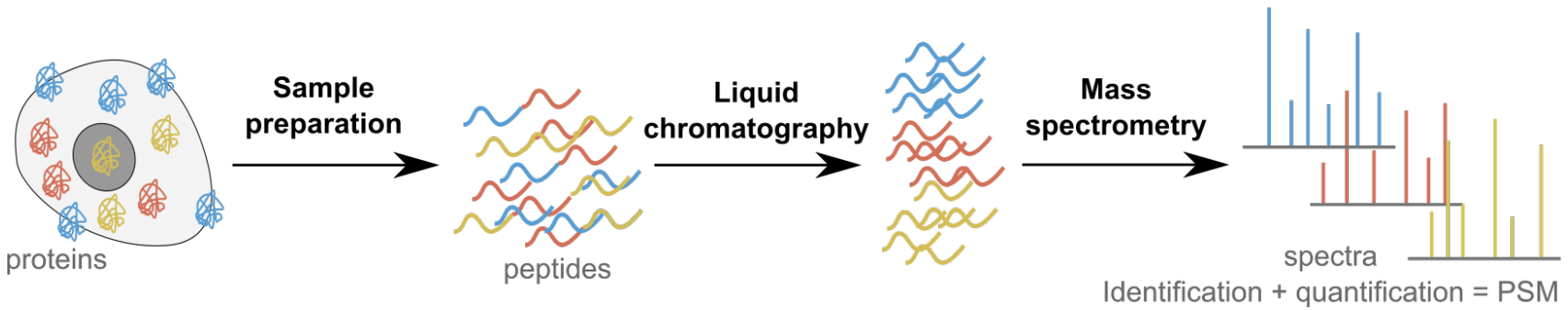




Source: Franks et al. (2017), Specht et al. (2020).

Single-cell proteomics

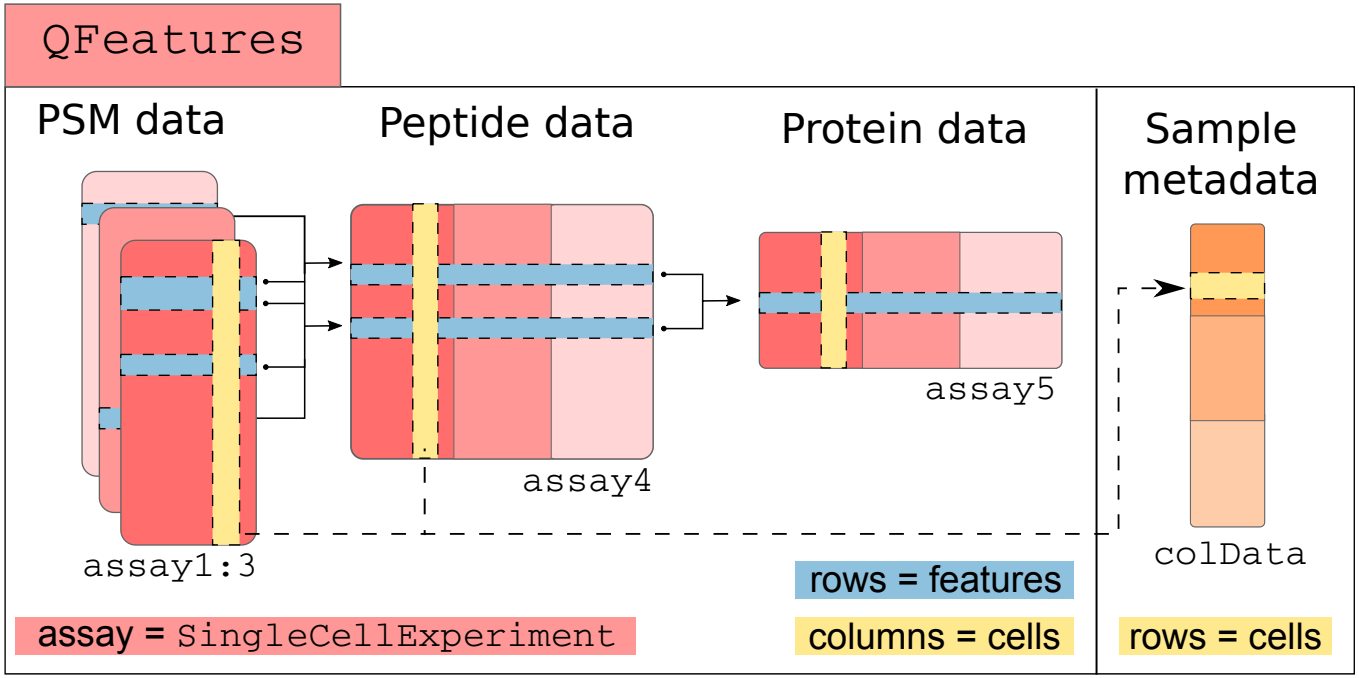
Single-cell proteomics recently achieved a milestone by quantifying $>$ 1000 proteins for $>$ 1000 single cells (Specht et al. (2020)).



- Label-free quantification: accurate quantification, but low throughput and low identification rate
- Multiplexed: label cross contamination, but high throughput and increased identification rate

Our contribution

We offer a solution to the lack of good computational tools for handling SCP data.



- `scpdata` disseminates curated SCP data sets for method development and benchmarking
- `scp` implements functions to streamline the analysis of SCP data

SCP pipeline

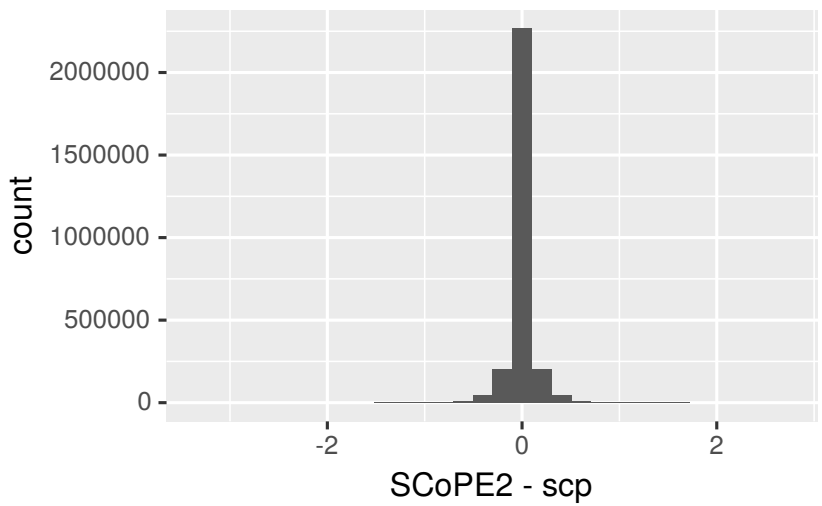
1. Input data
2. QC on features
3. QC on samples
4. Peptide aggregation
5. Log-normalization
6. Feature selection
7. Imputation
8. Protein aggregation
9. Data integration
10. Dimension reduction

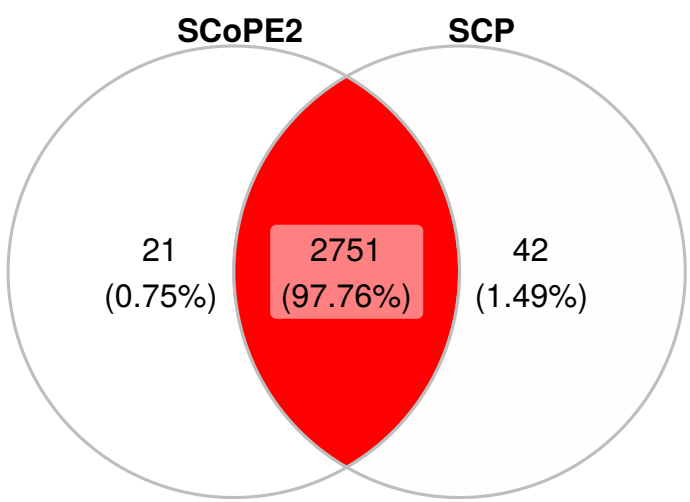
```
readSCP(quantTable = quantData,
        metaTable = metaData,
        channelCol = "Channel",
        batchCol = "Set") %>%
  zeroIsNA(i = 1:4) %>%
  filterFeatures(~ Potential.contaminant != "+") %>%
  computeSCR(i = 1:4,
             colDataCol = "SampleType",
             carrierPattern = "Carrier",
             samplePattern = "Monocyte") %>%
  filterFeatures(~ .meanSCR < 0.1) %>%
  subsetByAssay(dims(.)[1, ] > 150) %>%
  computeMedianCV(i = 1:3,
                  proteinCol = "protein",
                  peptideCol = "peptide") %>%
  aggregateFeaturesOverAssays(i = 1:3,
                              name = 4:6,
                              fcol = "peptide",
                              fun = robustSummary) %>%
  joinAssays(i = 4:6, name = "peptides") %>%
  normalize(i = "peptides",
            method = "median", na.rm = TRUE) %>%
  logTransform(i = "normAssay",
               base = 2) %>%
  impute(i = "normAssay",
         method = "knn") %>%
  aggregateFeatures(i = "logAssay",
                    name = "proteins",
                    fcol = "protein") ->
  scp
```

The SCoPE2 dataset I

SCoPE2 dataset (Specht et al. (2020)) = current state-of-the-art SCP dataset

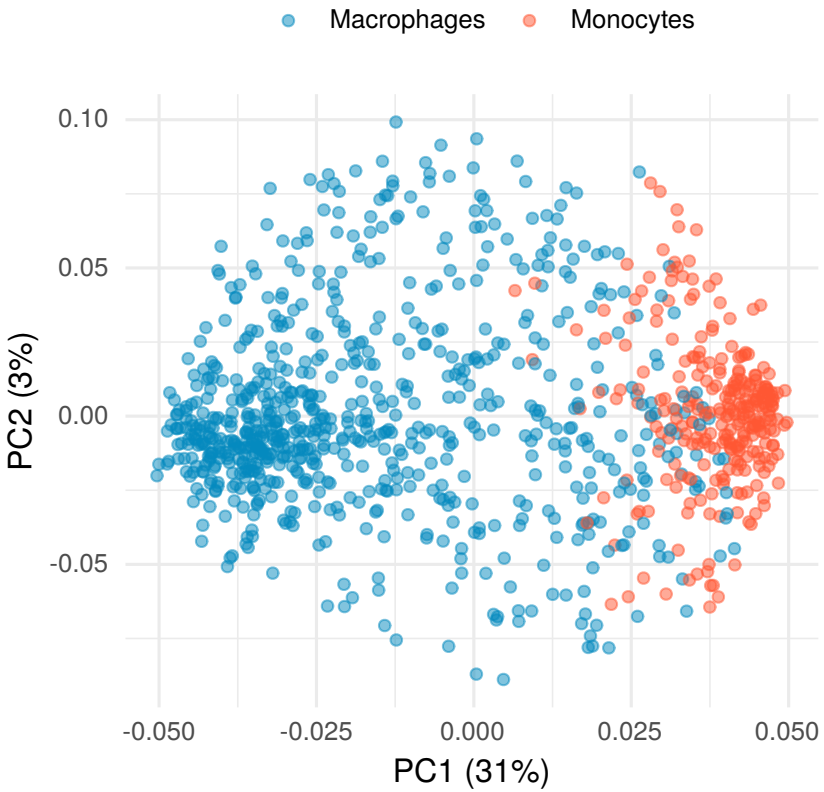
Replication of the analysis using `scp`:



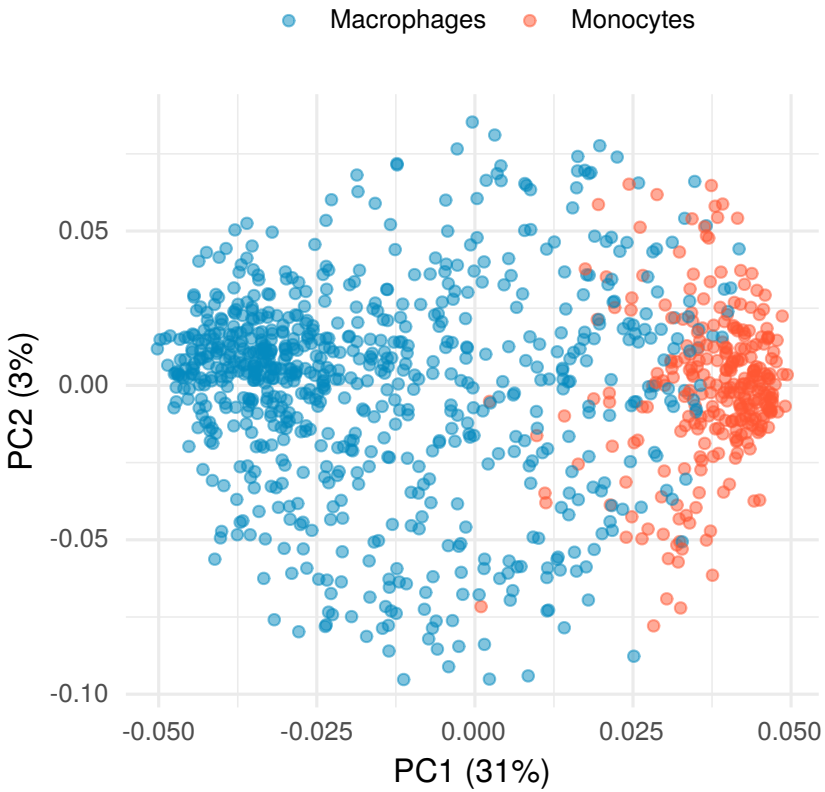


The SCoPE2 dataset II

SCoPE2



`scp`



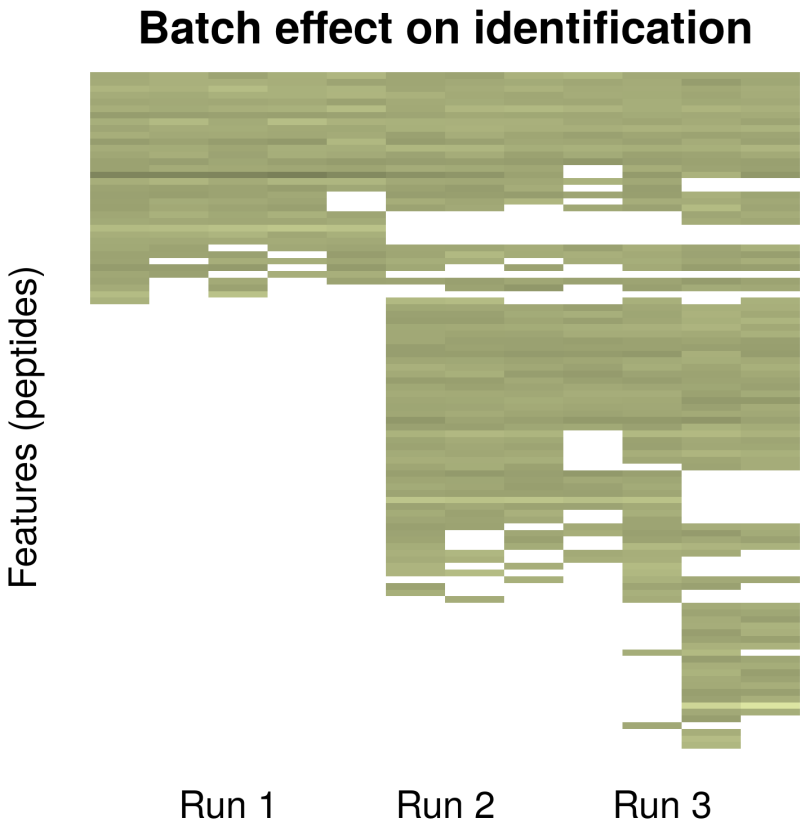
Replication: conclusion

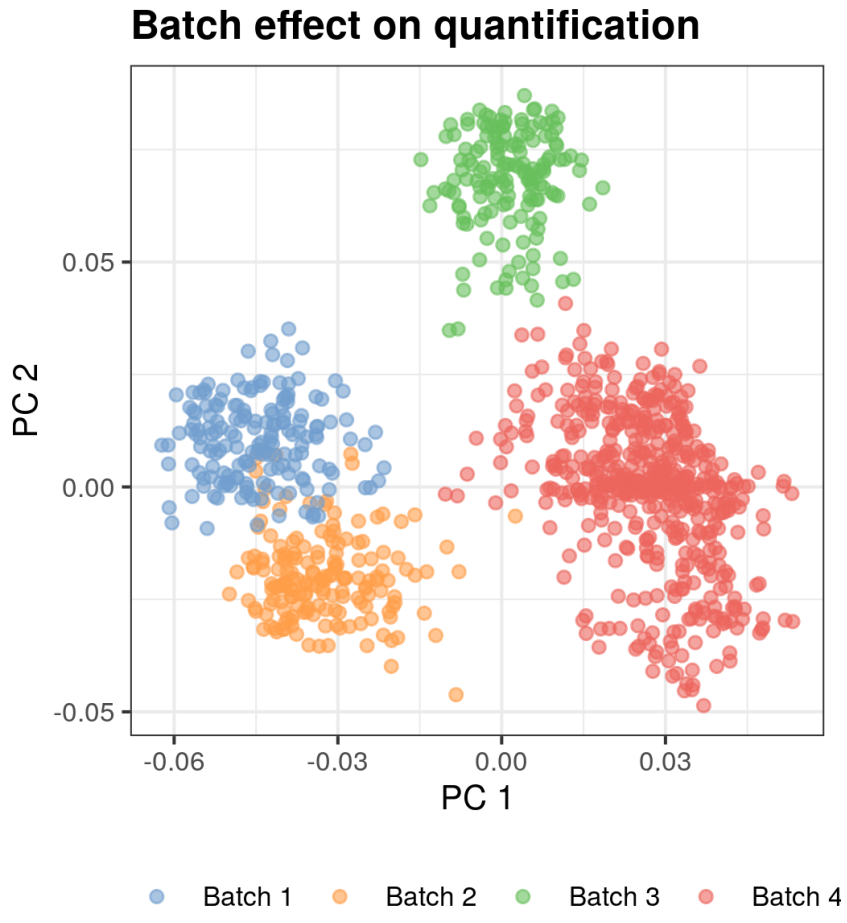
`scp` provides a standardized pipeline for unified and reproducible analysis of SCP data:

1. SCoPE2 (Specht et al. (2020)): almost perfect replication, new metrics included in `scp`, highlighted issues and possible improvements
2. Trajectory analysis on chicken utricle (Zhu et al. (2019)): lack of good documentation

This demonstrate the successful application of our software to various SCP datasets.

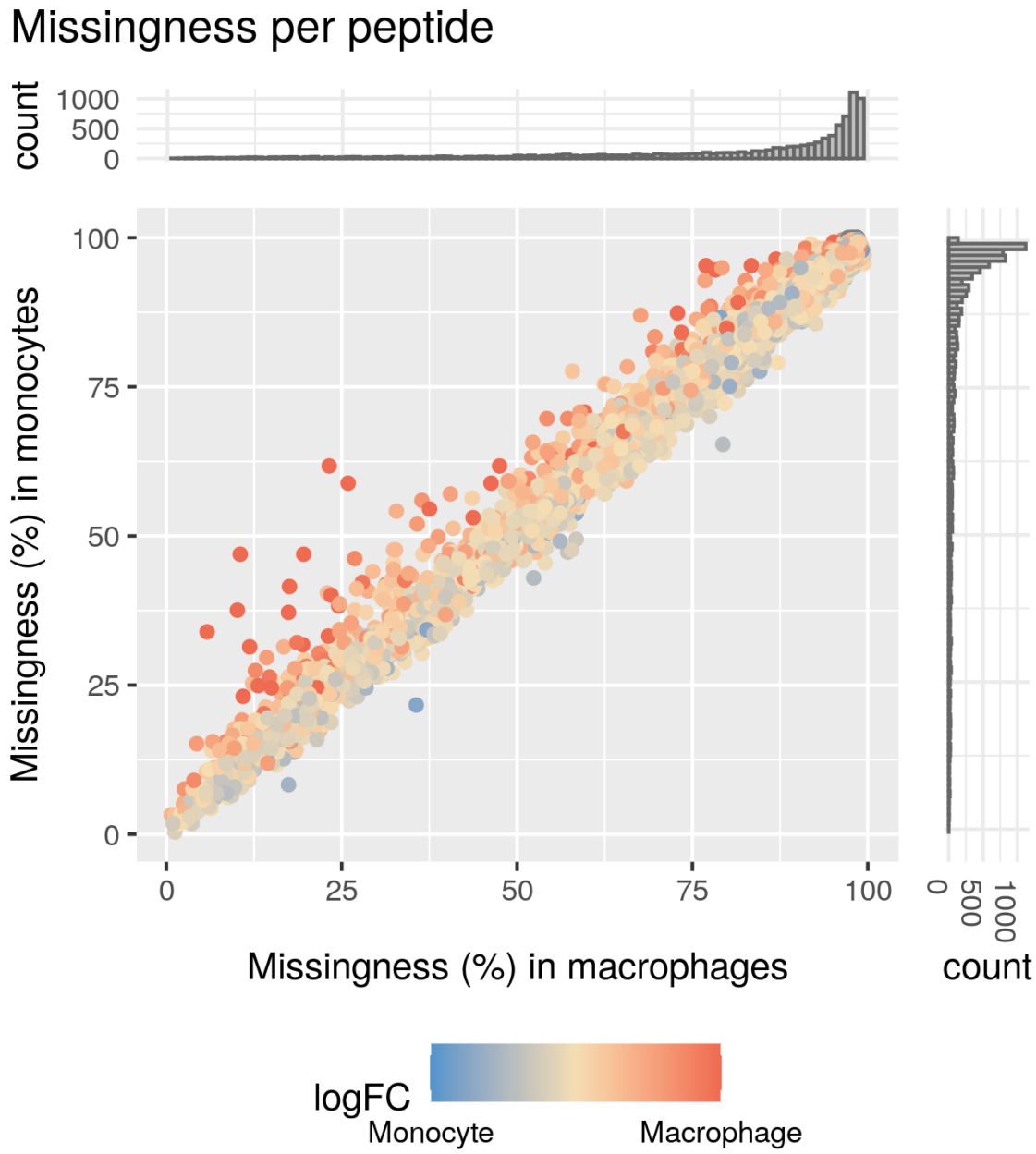
SCP challenges: batch effect





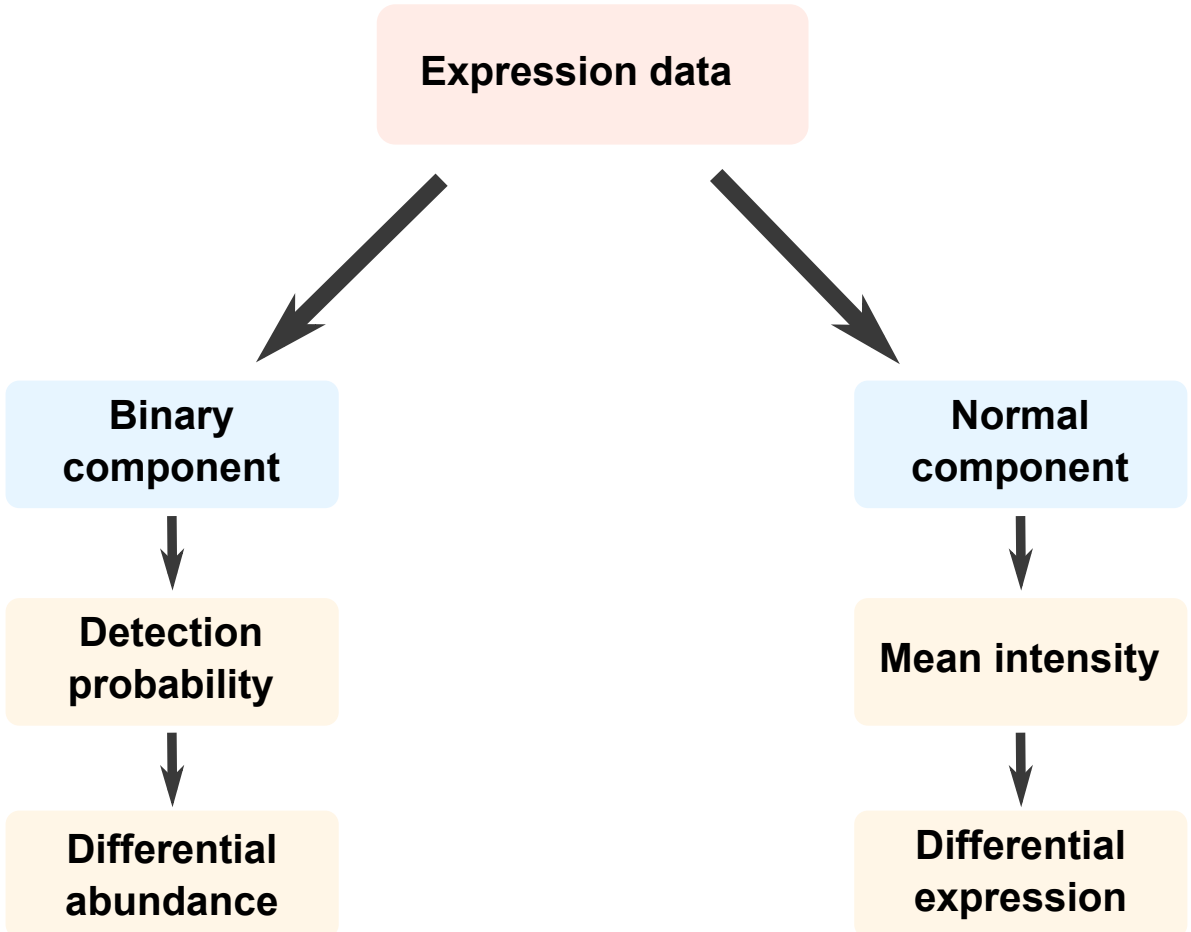
SCP challenges: missingness

- Biological missingness
- Technical missingness
- Both



SCP challenges: data modeling

Hurdle model (Goeminne et al. (2020))



Takehome message

- SCP is an emerging but very promosing field!
- We developed a computational infrastructure to formalize SCP data analyses
- The infrastructure could be applied to reproduce 2 published analyses
- Exciting challenges are yet to be solved

Acknowledgements

Many thanks to my promoter **Pr. Laurent Gatto**

Thanks you for your attention!

I'm happy to take questions now or at the discussion tables

See you at the EuroBioc2020 (online)

UCLouvain

fnrs
LA LIBERTÉ DE CHERCHER

References I

Alexander Franks, Edoardo Airoldi, and Nikolai Slavov. Post-transcriptional regulation across human tissues. *PLoS Comput. Biol.*, 13(5):e1005535, May 2017.

Ludger J E Goeminne, Adriaan Sticker, Lennart Martens, Kris Gevaert, and Lieven Clement. MSqRob takes the missing hurdle: Uniting intensity- and Count-Based proteomics. *Anal. Chem.*, 92(9):6278–6287, May 2020.

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. October 2020.

Ying Zhu, Mirko Scheibinger, Daniel Christian Ellwanger, Jocelyn F Krey, Dongseok Choi, Ryan T Kelly, Stefan Heller, and Peter G Barr-Gillespie. Single-cell proteomics reveals changes in expression during hair-cell development. *Elife*, 8, November 2019.