

Standardised and reproducible analysis of mass spectrometry-based single-cell proteomics data

Laurent Gatto, Christophe Vanderaa

CBIO, de Duve Institute, UCLouvain

18 August 2020

Outline

Introduction

Data framework

scp package

scp showcase

Replication results

Conclusion

MS-SCP: Mass spectrometry-based single-cell proteomics

MS-SCP consist of shotgun proteomics at single-cell level

- ▶ SCoPE2 quantifies thousands of proteins x thousands single-cells
- ▶ Full protocole available
- ▶ Full analysis script available

BUT

Lack of standardized analysis software

Provide a suite of software package dedicated to MS-SCP that fulfill:

- ▶ User-friendly
- ▶ Computationaly efficient
- ▶ Modularity: integrate other software packages
- ▶ Promote reproducibility
- ▶ Platform-independent
- ▶ Free of charge

R/Bioconductor is an ideal environment

Outline

Introduction

Data framework

scp package

scp showcase

Replication results

Conclusion

scpdata: distributes published MS-SCP datasets (e.g. SCoPE2 dataset)

scp: provides functionality for manipulating the MS-SCP data structure

Outline

Introduction

Data framework

scp package

scp showcase

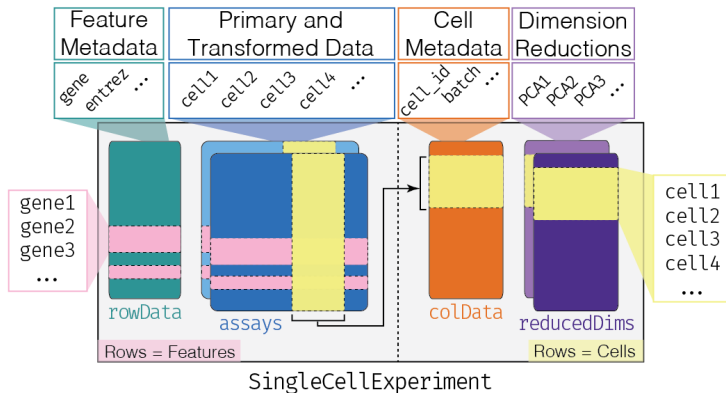
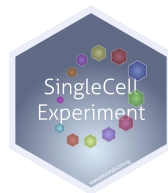
Replication results

Conclusion

Data infrastructure (1)

scp package

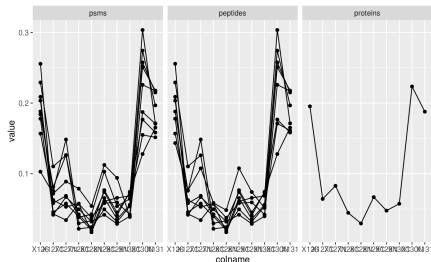
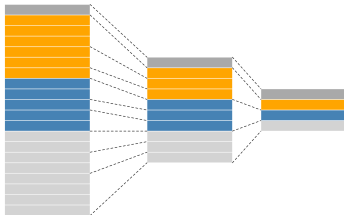
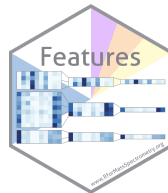
SingleCellExperiment: provides dedicated framework for single-cell data analysis.
Available on Bioconductor.



Data infrastructure (2)

scp package

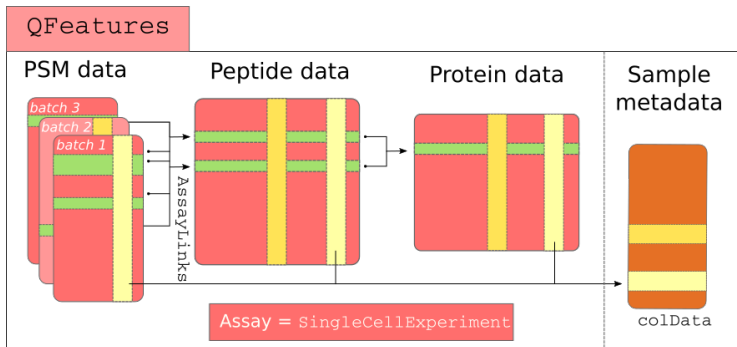
QFeatures: data framework dedicated to manipulate and process MS-based quantitative data.
Submitted to Bioconductor.



Data infrastructure (3)

scp package

`scp` = `SingleCellExperiment` + `QFeatures`



Load the SCoPE2 dataset called `specht2019v2`

```
1 library(scpdata)
2 data("specht2019v2")
```

Dataset overview

```
1 show(specht2019v2)
```

```
An instance of class QFeatures containing 179 assays:
[1] 190222S_LCA9_X_FP94AA: SingleCellExperiment with 2823 rows and 11 col...
[2] 190222S_LCA9_X_FP94AB: SingleCellExperiment with 4297 rows and 11 col...
[3] 190222S_LCA9_X_FP94AC: SingleCellExperiment with 4956 rows and 11 col...
...
[177] 191110S_LCB7_X_APNOV16plex2_Set_9: SingleCellExperiment with 4626 r...
[178] peptides: SingleCellExperiment with 9208 rows and 1018 columns
[179] proteins: SingleCellExperiment with 2772 rows and 1018 columns
```

(test slide: to discuss)

The sample metadata can be retrieved in the `colData`

```
1 colData(specht2019v2)
```

DataFrame with 2517 rows and 6 columns

sortday	digest	Set	Channel	SampleType	lcbatch
190222S_LCA9_X_FP94AA_RI1		<character>	<character>	<character>	<character>
s8	N	190222S_LC...	RI1	Carrier	LCA9
190222S_LCA9_X_FP94AA_RI2		190222S_LC...	RI2	Reference	LCA9
s8	N				
190222S_LCA9_X_FP94AA_RI3		190222S_LC...	RI3	Unused	LCA9
s8	N				
190222S_LCA9_X_FP94AA_RI4		190222S_LC...	RI4	Macrophage	LCA9
s8	N				
190222S_LCA9_X_FP94AA_RI5		190222S_LC...	RI5	Macrophage	LCA9
s8	N				
...	
191110S_LCB7_X_APNOV16plex2_Set_9_RI12		191110S_LC...	RI12	Macrophage	LCB7
s9	U				
191110S_LCB7_X_APNOV16plex2_Set_9_RI13		191110S_LC...	RI13	Macrophage	LCB7
s9	U				
191110S_LCB7_X_APNOV16plex2_Set_9_RI14		191110S_LC...	RI14	Macrophage	LCB7
s9	U				
191110S_LCB7_X_APNOV16plex2_Set_9_RI15		191110S_LC...	RI15	Monocyte	LCB7
s9	U				
191110S_LCB7_X_APNOV16plex2_Set_9_RI16		191110S_LC...	RI16	Macrophage	LCB7
s9	U				

The sample metadata can be retrieved in the `colData`

```
1 colData(specht2019v2)
```

- ▶ Batch name
- ▶ Channel name
- ▶ Sample info: sample type, treatment, ...
- ▶ Batch info: chromatographic batch, digestion batch, ...

The sample metadata can be retrieved in the `colData`

```
1 colData(specht2019v2)
```

- ▶ Batch name
- ▶ Channel name
- ▶ Sample info: sample type, treatment, ...
- ▶ Batch info: chromatographic batch, digestion batch, ...

The feature metadata can be retrieved in the `rowData`, but assay specific

```
1 rowData(specht2019v2[[1]])
```

- ▶ PSM level: reverse hit, PEP, m/z value, charge, ...
- ▶ Peptide level: sequence, length, modification, mass, ...
- ▶ Protein level: name, sequence, gene name, ...

1. Load data

PSM data

```
[1] 190222S_LCA9_X_FP94AA: SingleCellExperiment with 2823 rows and 11 columns
[2] 190222S_LCA9_X_FP94AB: SingleCellExperiment with 4297 rows and 11 columns
[3] 190222S_LCA9_X_FP94AC: SingleCellExperiment with 4956 rows and 11 columns
...
[177] 191110S_LCB7_X_APNOV16plex2_Set_9: SingleCellExperiment with 4626 rows and 16 columns
```

1. Load data

PSM data

```
[1] 190222S_LCA9_X_FP94AA: SingleCellExperiment with 2823 rows and 11 columns  
[2] 190222S_LCA9_X_FP94AB: SingleCellExperiment with 4297 rows and 11 columns  
[3] 190222S_LCA9_X_FP94AC: SingleCellExperiment with 4956 rows and 11 columns  
...  
[177] 191110S_LCB7_X_APNOV16plex2_Set_9: SingleCellExperiment with 4626 rows and 16 columns
```

2. PSM filtering
3. Expression channel by reference channel division
4. PSM to peptides aggregating
5. Single cells filtering based on median CV
6. Normalization
7. Removal of highly missing peptides
8. Log-transformation

Peptide data

```
[178] peptides: SingleCellExperiment with 9208 rows and 1018 columns
```


1. Load data

PSM data

```
[1] 190222S_LCA9_X_FP94AA: SingleCellExperiment with 2823 rows and 11 columns
[2] 190222S_LCA9_X_FP94AB: SingleCellExperiment with 4297 rows and 11 columns
[3] 190222S_LCA9_X_FP94AC: SingleCellExperiment with 4956 rows and 11 columns
...
[177] 191110S_LCB7_X_APNOV16plex2_Set_9: SingleCellExperiment with 4626 rows and 16 columns
```

2. PSM filtering
3. Expression channel by reference channel division
4. PSM to peptides aggregating
5. Single cells filtering based on median CV
6. Normalization
7. Removal of highly missing peptides
8. Log-transformation

Peptide data

```
[178] peptides: SingleCellExperiment with 9208 rows and 1018 columns
```

9. Peptides to proteins aggregation
10. Normalization
11. Imputation
12. Batch correction

Protein data

```
[179] proteins: SingleCellExperiment with 2772 rows and 1018 columns
```

Outline

Introduction

Data framework

scp package

scp showcase

Replication results

Conclusion

Filter out features based on the feature metadata

Example: filter out reverse hits. The filter is applied to the `Reverse` field in the feature metadata

```
1 filterFeatures(specht2019v2 ,  
2               ~ Reverse != "+")
```

Source code in `QFeatures`

Interesting metrics for MS-SCP quality control:

- ▶ Sample to carrier ratio: ratio of the carrier channel intensity signal over the sample channel intensity
- ▶ Peptide FDR¹: expected rate of wrongly assigned features to a given peptide
- ▶ Cell median CV²: reliability of the protein quantification summarized over each cell.

Example:

```
1 computeMedianCV(specht2019v2,  
2                 i = "peptides",  
3                 proteinCol = "protein",  
4                 peptideCol = "peptide",  
5                 batchCol = "Set")
```

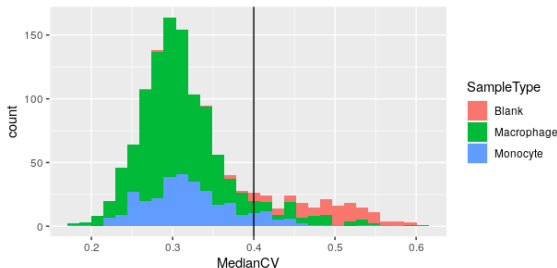
Source code in `scp`

¹false discovery rate

²coefficient of variation

QC metrics are stored in the data set for plotting or subsetting

```
1 library(tidyverse)
2 specht2019v2[["peptides"]] %>%
3   colData %>%
4   data.frame %>%
5   ggplot(aes(x = MedianCV,
6             fill = SampleType)) +
7   geom_histogram() +
8   geom_vline(xintercept = 0.4)
```



Feature aggregation includes 2 steps:

- ▶ Combine the quantitative data from multiple features to a single aggregated features
- ▶ Store the relationship between the parent features and the aggregated features

Example: aggregate peptides to proteins

```
1 aggregateFeatures(specht2019v2 ,  
2                   i = "peptides",  
3                   name = "proteins",  
4                   fcol = "protein",  
5                   fun = colMedians, na.rm = TRUE)
```

Source code in `QFeatures`

0's can be either **biological** or **technical** zero. They are better related by NA's.

```
1 zeroIsNA(specht2019v2,  
2         i = "peptides")
```

Features containing too many missing data (e.g. $\geq 99\%$) should be removed

```
1 filterNA(specht2019v2,  
2         i = "peptides",  
3         pNA = 0.99)
```

Source code in `QFeatures`

Common data transformation can easily be applied:

- ▶ Normalization
- ▶ Log-transformation
- ▶ Imputation

Example: \log_2 -transformation:

```
1 logTransform(specht2019v2 ,  
2             i = "peptides",  
3             base = 2,  
4             name = "peptides_log")
```

Source code in `QFeatures`

Some custom function can be applied to the data set too.

Example: batch correction using `sva::ComBat`. First, extract the data

```
1 sce <- specht2019v2[["proteins"]]
```

Build the correction matrix and apply the ComBat algorithm

```
1 batch <- colData(sce)$Set
2 model <- model.matrix(~ SampleType, data = colData(sce))
3 assay(sce) <- ComBat(dat = assay(sce),
4                       batch = batch,
5                       mod = model)
```

Add the corrected protein to the dataset and keep feature relationships

```
1 addAssay(specht2019v2,
2          sce,
3          name = "proteins_batchC") %>%
4 addAssayLinkOneToOne(from = "proteins",
5                       to = "proteins_batchC")
```

Outline

Introduction

Data framework

scp package

scp showcase

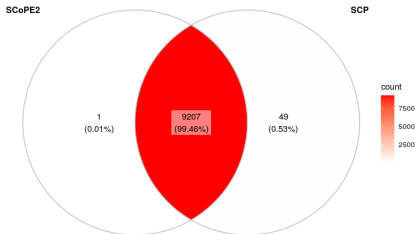
Replication results

Conclusion

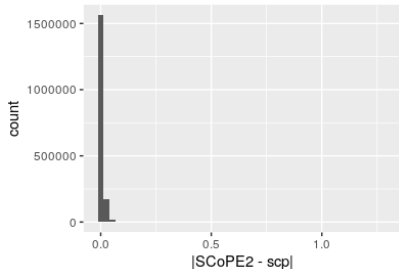
Benchmarking data matrices

Replication results

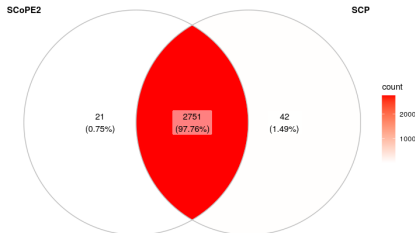
Peptides



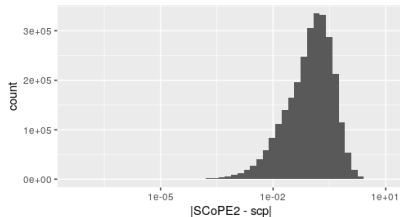
Benchmark of the peptide data



Proteins



Benchmark of the protein data



Replicate figures from SCoPE2 (1)

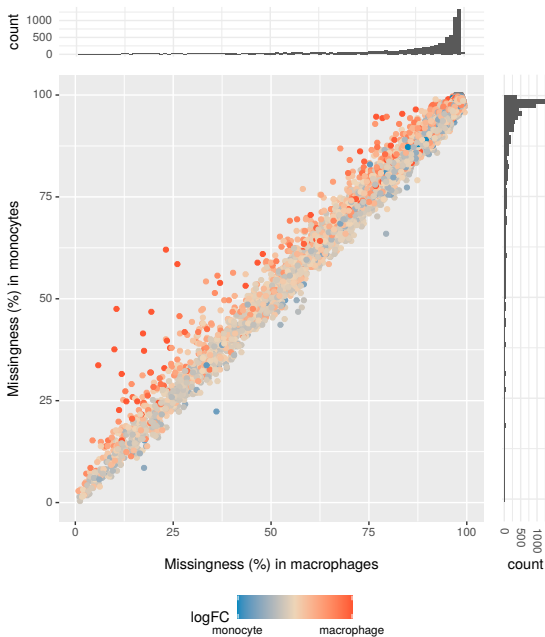
Replication results

Replicate figures from SCoPE2 (2)

Replication results

Missingness

Replication results



Outline

Introduction

Data framework

scp package

scp showcase

Replication results

Conclusion

- ▶ `scp` package suite provides a standardized environment for performing MS-SCP data analysis
- ▶ Flexibly reproduce existing analyses from different groups or protocols (multiplex vs label free)

Advantages:

- ▶ Allow automation of the analysis
- ▶ Facilitate new computational developments
- ▶ Promotes reproducibility
- ▶ Increases field visibility
- ▶ Include other modalities: scRNA-Seq, ATAC-Seq, etc

Packages

- ▶ `scp`: GitHub repository `UClouvain-CBIO/scp`
- ▶ `scpdata`: coming soon
- ▶ `QFeatures`: GitHub repository
`rformassspectrometry/QFeatures`
- ▶ `SingleCellExperiment`: Bioconductor

SCoPE2 reproduction vignette

Available at...

Slides and source code

Available at...

