Data preprocessing in R

2025-06-01

Contents

Data pre-processing

SCTransform

Load needed libraries:

```
library(Seurat)
```

```
## Loading required package: SeuratObject
```

```
## Loading required package: sp
```

```
##
## Attaching package: 'SeuratObject'
```

```
## The following objects are masked from 'package:base':
##
##     intersect, t
```

```
library(ggplot2)
library(sctransform)
library(Matrix)
library(reticulate)
library(scCustomize)
```

```
## scCustomize v3.0.1
## If you find the scCustomize useful please cite.
## See 'samuel-marsh.github.io/scCustomize/articles/FAQ.html' for citation info.
```

```
library(future)
options(future.globals.maxSize = 8000 * 1024^2)
```

Load data and create meta data dataframe:

```
setwd("/Users/sabrina/Library/CloudStorage/OneDrive-UniversityofCopenhagen/Thesis/CCC inference/RNA_data
data <- Read10X(data.dir = "/Users/sabrina/Library/CloudStorage/OneDrive-UniversityofCopenhagen/Thesis/C
counts_file <- "matrix.mtx"
labels_file <- "labels_subset.csv"
genes <- read.table("genes.tsv", header = FALSE, col.names = "gene_id")
dim(genes)
```

```
## [1] 25419     1
```

```
counts <- readMM(counts_file)
labels_df <- read.csv(labels_file)
```

```
cell_id_col <- "cell_id"
label_col <- "cell_type"
meta_df <- data.frame(cell_type = labels_df[[label_col]], row.names = labels_df[[cell_id_col]])
```

Create seurat object and run SCTransform:

```
data_so <- CreateSeuratObject(counts = data, meta.data = meta_df)
data_so <- SCTransform(data_so, verbose = TRUE)
```

```
## Running SCTransform on assay: RNA
```

```
## Warning: The `slot` argument of `GetAssayData()` is deprecated as of SeuratObject 5.0.0.
## i Please use the `layer` argument instead.
## i The deprecated feature was likely used in the Seurat package.
##   Please report the issue at <https://github.com/satijalab/seurat/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## vst.flavor='v2' set. Using model with fixed slope and excluding poisson genes.
```

```
## Calculating cell attributes from input UMI matrix: log_umi
```

```
## Variance stabilizing transformation of count matrix of size 19377 by 3126
```

```
## Model formula is y ~ log_umi
```

```
## Get Negative Binomial regression parameters per gene
```

```
## Using 2000 genes, 3126 cells
```

```
## Found 6 outliers - those will be ignored in fitting/regularization step
```

```
## Second step: Get residuals using fitted parameters for 19377 genes
```

```
## Computing corrected count matrix for 19377 genes
```

```
## Calculating gene attributes
```

```
## Wall clock passed: Time difference of 7.134672 secs
```

```
## Determine variable features
```

```
## Centering data matrix
```

```
## Place corrected count matrix in counts slot
```

```
## Warning: The `slot` argument of `SetAssayData()` is deprecated as of SeuratObject 5.0.0.
## i Please use the `layer` argument instead.
## i The deprecated feature was likely used in the Seurat package.
##   Please report the issue at <https://github.com/satijalab/seurat/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Set default assay to SCT
```

Check how many genes were filtered (non-variable) after SCTransform:

```
dim(data)
```

```
## [1] 25419  3126
```

```r
dim(data_so) # after SCTransform
```

```
## [1] 19377  3126
```

Keep matrix in .mtx format just as a precaution:

```r
sparse.data <- Matrix(data_so@assays$SCT$data, sparse = T)
# head(sparse.data)
writeMM(obj = sparse.data, file="/Users/sabrina/Library/CloudStorage/OneDrive-UniversityofCopenhagen/Th
```

```
## NULL
```

Anndata file creation

To be used to run LIANA+ on python. Activate conda environment and create anndata file:

```r
use_condaenv("/Users/sabrina/miniconda3/envs/liana_env", required = TRUE)
as.anndata(x = data_so, file_path = "/Users/sabrina/Library/CloudStorage/OneDrive-UniversityofCopenhage
```

```
## * Checking Seurat object validity
## * Extracting Data from SCT assay to transfer to anndata.
## The following columns were removed as they contain identical values for all
## rows:
## i orig.ident
## * Creating anndata object.
## * Writing anndata file:
##   "/Users/sabrina/Library/CloudStorage/OneDrive-UniversityofCopenhagen/Thesis/CCC
##   inference/Liana/rna_data.h5ad"
```

```
## AnnData object with n_obs × n_vars = 3126 × 19377
##     obs: 'nCount_RNA', 'nFeature_RNA', 'cell_type', 'nCount_SCT', 'nFeature_SCT'
##     var: 'names'
##     layers: 'data_SCT'
```

Cell-cell communication inference methods

CellChat

Load library and data:

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: igraph
```

```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:future':
##
##     %->%, %<-%

## The following object is masked from 'package:Seurat':
##
##     components

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

Create CellChat object:

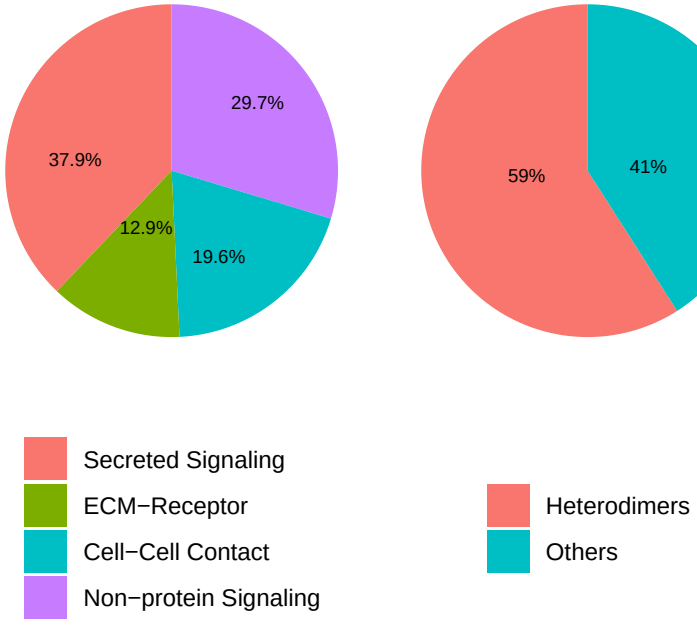
```
#cellChat <- createCellChat(object = seurat.obj, group.by = "ident", assay = "RNA")
x <- createCellChat(object = data_so.input, meta = meta_df, group.by = "cell_type") # IMPORTANT: do not
```

```
## [1] "Create a CellChat object from a data matrix"

## Warning in createCellChat(object = data_so.input, meta = meta_df, group.by = "cell_type"): The 'meta

## Set cell identities for the new CellChat object
## The cell groups used for CellChat analysis are  DI6, MN, V0d, V0v, V1-Foxp2, V1-Pou6f2, V1-Rensh, V1
```

Select the correct database to use (in our case we use the entire mouse database to make sure it includes "Non-protein Signaling" i.e., metabolic and synaptic signaling).




```
## Rows: 3,379
## Columns: 28
## $ interaction_name         <chr> "TGFB1_TGFBR1_TGFBR2", "TGFB2_TGFBR1_TGFBR2",~
## $ pathway_name             <chr> "TGFb", "TGFb", "TGFb", "TGFb", "TGFb", "TGFb~
## $ ligand                   <chr> "Tgfb1", "Tgfb2", "Tgfb3", "Tgfb1", "Tgfb1", ~
## $ receptor                 <chr> "TGFbR1_R2", "TGFbR1_R2", "TGFbR1_R2", "ACVR1~
## $ agonist                  <chr> "TGFb agonist", "TGFb agonist", "TGFb agonist~
## $ antagonist               <chr> "TGFb antagonist", "TGFb antagonist", "TGFb a~
## $ co_A_receptor            <chr> "", "", "", "", "", "", "", "", "", "", "", "~
```

```
## $ co_I_receptor            <chr> "TGFb inhibition receptor", "TGFb inhibition ~
## $ annotation               <chr> "Secreted Signaling", "Secreted Signaling", "~
## $ interaction_name_2       <chr> "Tgfb1 - (Tgfbr1+Tgfbr2)", "Tgfb2  - (Tgfbr1+~
## $ evidence                 <chr> "KEGG: mmu04350", "KEGG: mmu04350", "KEGG: mm~
## $ is_neurotransmitter      <chr> "FALSE", "FALSE", "FALSE", "FALSE", "FALSE", ~
## $ ligand.symbol            <chr> "Tgfb1", "Tgfb2", "Tgfb3", "Tgfb1", "Tgfb1", ~
## $ ligand.family            <chr> "TGF-beta", "TGF-beta", "TGF-beta", "TGF-beta~
## $ ligand.location          <chr> "Extracellular matrix, Secreted, Extracellula~
## $ ligand.keyword           <chr> "Disease variant, Signal, Reference proteome,~
## $ ligand.secreted_type     <chr> "growth factor", "growth factor", "cytokine;g~
## $ ligand.transmembrane     <chr> "FALSE", "FALSE", "TRUE", "FALSE", "FALSE", "~
## $ receptor.symbol          <chr> "Tgfbr1, Tgfbr2", "Tgfbr1, Tgfbr2", "Tgfbr1, ~
## $ receptor.family          <chr> "Protein kinase superfamily, TKL Ser/Thr prot~
## $ receptor.location        <chr> "Cell membrane, Secreted, Membrane raft, Cell~
## $ receptor.keyword         <chr> "Membrane, Secreted, Disulfide bond, Kinase, ~
## $ receptor.surfaceome_main <chr> "Receptors", "Receptors", "Receptors", "Recep~
## $ receptor.surfaceome_sub  <chr> "Act.TGFB;Kinase", "Act.TGFB;Kinase", "Act.TG~
## $ receptor.adhesome        <chr> "", "", "", "", "", "", "", "", "", "", "", "~
## $ receptor.secreted_type   <chr> "", "", "", "", "", "", "", "", "", "", "", "~
## $ receptor.transmembrane   <chr> "TRUE", "TRUE", "TRUE", "TRUE", "TRUE", "TRUE~
## $ version                  <chr> "CellChatDB v1", "CellChatDB v1", "CellChatDB~
```

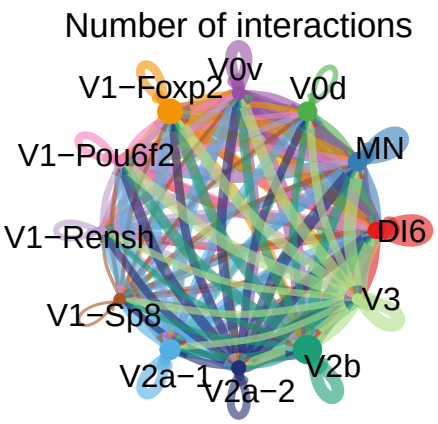
```
## The number of highly variable ligand-receptor pairs used for signaling inference is 1870
```

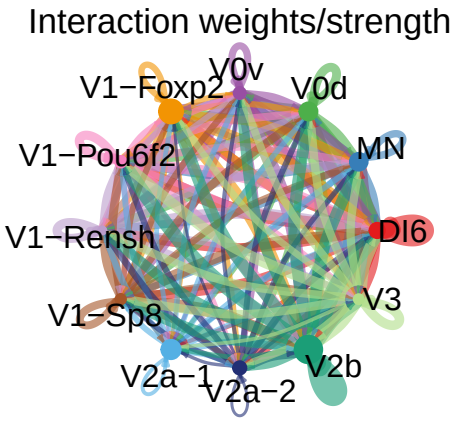
```
## [1] 48.43908
```

```
## triMean is used for calculating the average gene expression per cell group.
## [1] ">>> Run CellChat on sc/snRNA-seq data <<< [2025-08-15 10:12:29.684653]"
## [1] ">>> CellChat inference is done. Parameter values are stored in `object@options$parameter` <<< [1
```

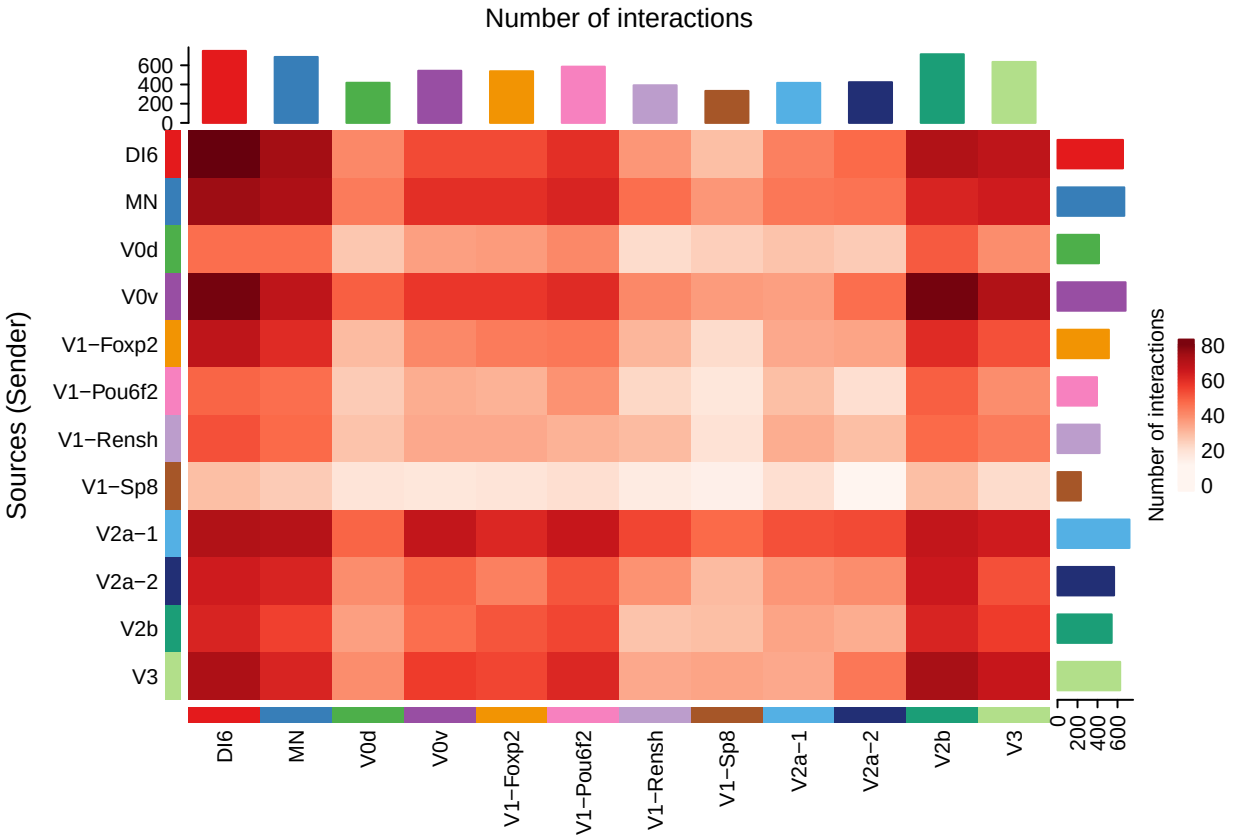
```
##         source target ligand receptor        prob pval interaction_name
## 6433  V1-Foxp2     V3  Cadm3  Epb41l1 0.008414724 0.00    CADM3_EPB41L1
## 6434 V1-Pou6f2     V3  Cadm3  Epb41l1 0.008414724 0.00    CADM3_EPB41L1
## 6435  V1-Rensh     V3  Cadm3  Epb41l1 0.008414724 0.04    CADM3_EPB41L1
## 6436     V2a-1     V3  Cadm3  Epb41l1 0.024826359 0.00    CADM3_EPB41L1
## 6437       V2b     V3  Cadm3  Epb41l1 0.008414724 0.00    CADM3_EPB41L1
## 6438        V3     V3  Cadm3  Epb41l1 0.008414724 0.00    CADM3_EPB41L1
##      interaction_name_2 pathway_name        annotation evidence
## 6433    Cadm3 - Epb41l1         CADM Cell-Cell Contact  uniprot
## 6434    Cadm3 - Epb41l1         CADM Cell-Cell Contact  uniprot
## 6435    Cadm3 - Epb41l1         CADM Cell-Cell Contact  uniprot
## 6436    Cadm3 - Epb41l1         CADM Cell-Cell Contact  uniprot
## 6437    Cadm3 - Epb41l1         CADM Cell-Cell Contact  uniprot
## 6438    Cadm3 - Epb41l1         CADM Cell-Cell Contact  uniprot
```

```
## [1] 56497.55
```

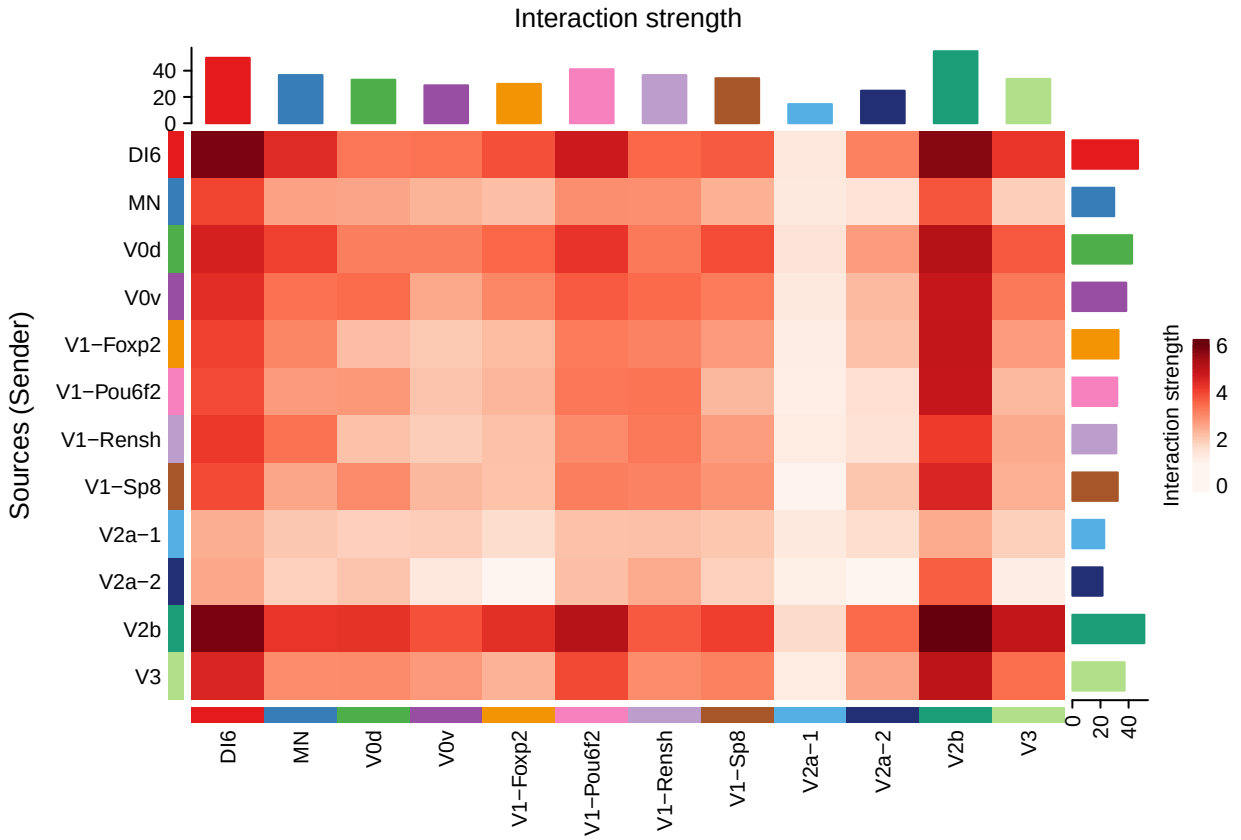





```
## Do heatmap based on a single object
```




```
## Do heatmap based on a single object
```




```
length(unique(x@DB$interaction$interaction_name))
```

```
## [1] 3379
```

```
length(unique(x@net$LRs))
```

```
## [1] 246
```