QUALITY CONTROL AND PREPROCESSING, SINGLE CELL EXPERIMENTS

Introduction

The script 'QC_plus_doubletsdet.R' located in this repository is valid for 10X-format raw counts. Single Cell RNA sequencing (scRNAseq) projects in our setting start with cell isolation from skeletal muscle and sorting, followed by Chromium transcriptome 3' protocol (10X Genomics) which includes barcoded gel beads library preparation, unique molecular identifier (UMI) linking, sequencing, alignment and Cell Ranger filtering. The 10X protocol is performed by a third party laboratory, then resulting raw count matrix and metadata are received for our downstream analysis. Even in this scenario, rigorous pre-processing is mandatory.

The present document is only for practical instructions regarding the QC script. Concepts and details about scRNAseq and quality control are found in INMG_SingleCell/scRNAseqMuscleNiche.pdf.

Getting started

IMPORTANT Use **one single** batch data when running this script, as doublet detection procedures should only be applied to libraries generated in the same experimental batch.

The input is the experiment in the form of a folder (here for our illustration 'dorsowt2'), containing the raw count matrix and its respective metadata, organized as follows:

- data/
  - dorsowt2/
    - barcodes.tsv.gz
    - features.tsv.gz
    - matrix.mtx.gz

Within the R code, change 'exper' variable accordingly, check also working directory ('prloc' variable). I recommend to stick to default location (HOME) for 'QC_single_cell'. Launch from RStudio, and if any difficulties are encountered check 'results/outputsfile.txt' to see at which level error occurs. An executable version will be available to be able to run into a bash loop, taking as argument the experiment's raw matrix folder name.

'QC_plus_doubletsdet.R' must be used if features are present as gene symbols ("Gsn"). One version for features in the form of Ensembl identifiers ("ENSMUSG00000026879") will be soon available.

Brief Illustration

Here we load an '_END.RData' already generated after running 'QC_plus_doubletsdet.R' on publicly available 10X data in mouse model from GEO (accession code GSM3614993). Lets see SingleCellExperiment object and dimensions:

```
load(paste0("rdatas/",exper,"_END.RData"))

sce
```

```
## class: SingleCellExperiment
## dim: 17616 2418
## metadata(0):
## assays(2): counts logcounts
## rownames(17616): Gm1992 Sox17 ... DHRSX CAAA01147332.1
## rowData names(8): genes_names ensembl.id ... mean detected
## colnames(2418): AAACCTGCAGGACCCT-1 AAACGGGAGCTGCGAA-1 ...
##   TTTGTCATCAATCACG-1 TTTGTCATCGAGGTAG-1
## colData names(17): n_mm_umi n_hg_umi ... is_cell doublet_score
## reducedDimNames(0):
## spikeNames(0):
## altExpNames(0):
```

We expect only M musculus but we found out gene symbols from H sapiens (this is negligeable anyway, and metrics did not show relevant contamination):

```
table(rowData(sce)$species)
```

```
##
## Homo sapiens Mus musculus
##           17        17599
```

```
tail(rowData(sce)[rowData(sce)$species %in% "Homo sapiens",])
```

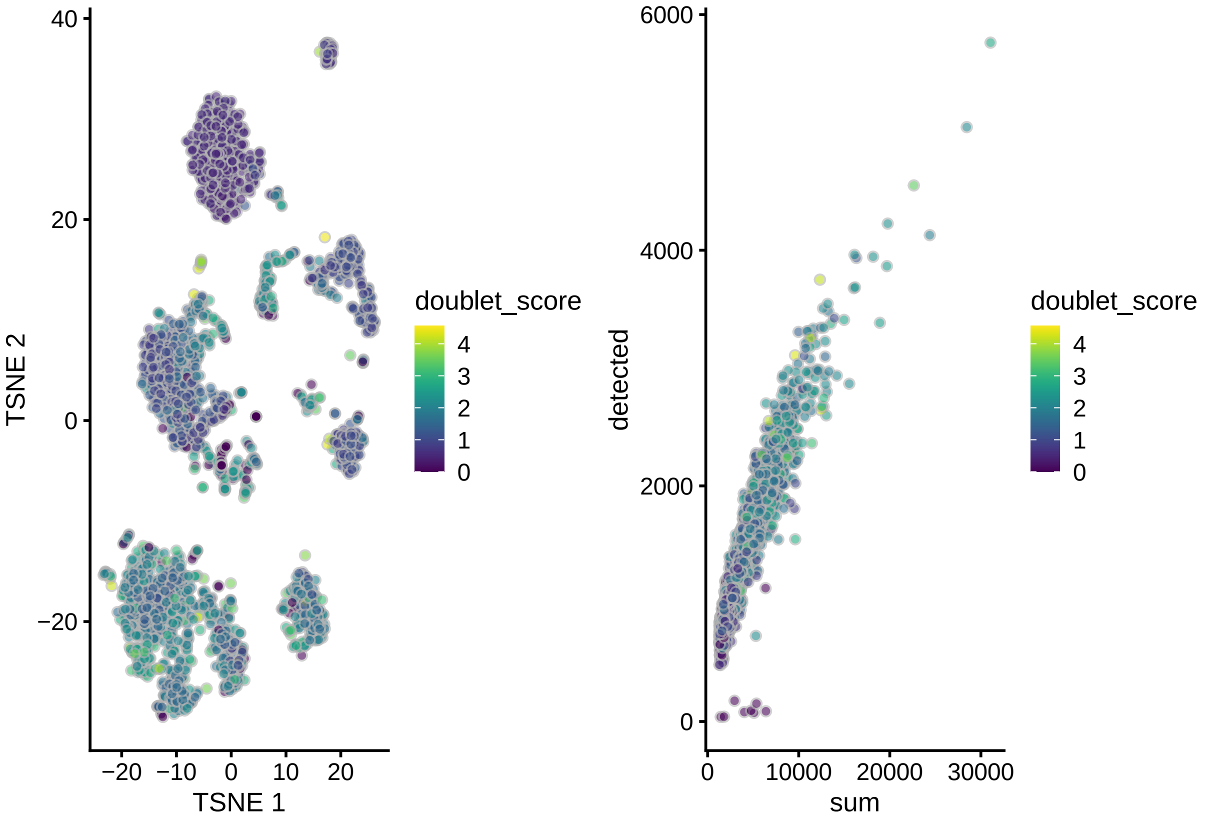
```
## DataFrame with 6 rows and 8 columns
##       genes_names                         ensembl.id     chr_pos is_genomic
##       <character>                        <character> <character>  <logical>
## C7             C7 ENSG00000112936_ENSMUSG00000079105                  FALSE
## WDR97       WDR97                    ENSG00000179698           8       TRUE
## C2             C2 ENSG00000166278_ENSMUSG00000024371                  FALSE
## C3             C3 ENSG00000125730_ENSMUSG00000024164                  FALSE
## PISD         PISD                    ENSG00000241878          22       TRUE
## DHRSX       DHRSX                    ENSG00000169084           X       TRUE
##            species expressed                 mean           detected
##        <character> <logical>            <numeric>          <numeric>
## C7    Homo sapiens      TRUE   0.0111019736842105  0.904605263157895
## WDR97 Homo sapiens      TRUE 0.000411184210526316 0.0411184210526316
## C2    Homo sapiens      TRUE   0.0826480263157895   4.35855263157895
## C3    Homo sapiens      TRUE     1.58059210526316   26.3569078947368
## PISD  Homo sapiens      TRUE    0.189555921052632           14.84375
## DHRSX Homo sapiens      TRUE   0.0875822368421053   7.93585526315789
```

Doublets detection

```
tsnepl <- plotTSNE(sce[rowData(sce)$expressed,sce$is_cell], colour_by="doublet_score")
```

```
## Warning: call 'runTSNE' explicitly to compute results
```

```
detfeat <- scater::plotColData(sce, x="sum",y="detected",colour_by="doublet_score")
tsnepl + detfeat
```



output All figures in .pdf format are saved in 'results/' whereas 'rdatas/__END.RData' file corresponds to filtered sce object.

Acknowledgements

Many thanks to Dr. L Modolo for most of this code

sources

I strongly encourage the user/developer to read:

http://perso.ens-lyon.fr/laurent.modolo/scRNA/

https://bioconductor.org/packages/release/bioc/vignettes/scater/inst/doc/overview.html

author Johanna Galvis-Lascroux, 2020