

# **A comprehensive single cell RNA analysis pipeline**

**Mingbo Cheng**

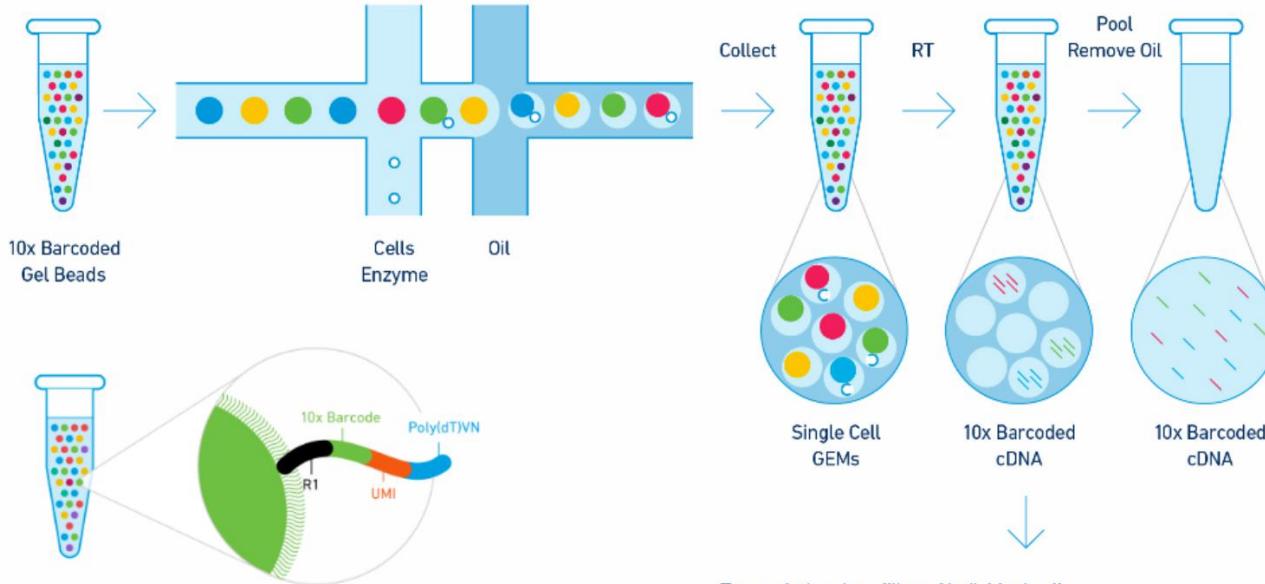
# Outline

- Overview of data analysis in scRNAseq data
- Single cell pipeline introduction
  - overview
  - output example
  - quick start
  - parameter settings
  - frequently asked questions

# Outline

- Overview of data analysis in scRNAseq data
- Single cell pipeline introduction
  - overview
  - output example
  - quick start
  - parameter settings
  - frequently asked questions

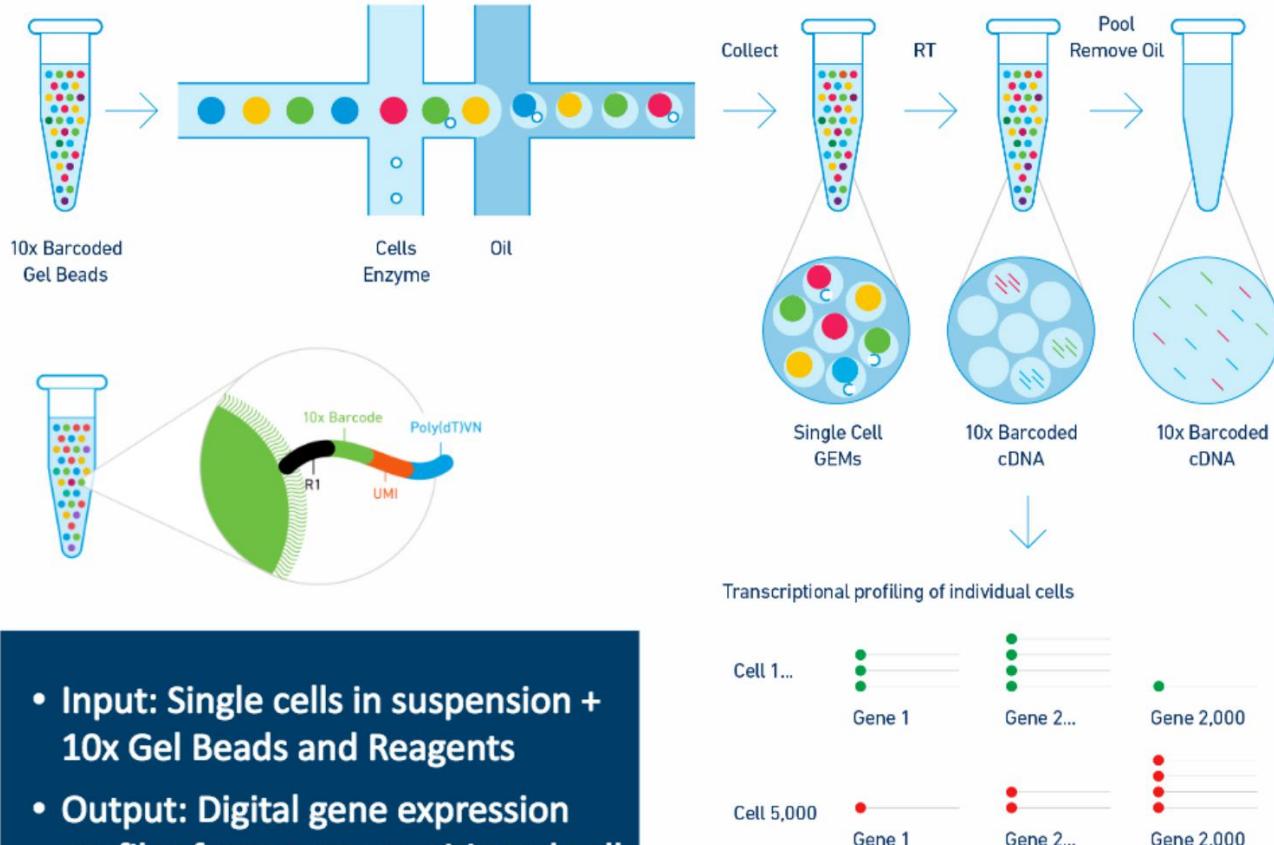
# 10X single cell RNA analysis



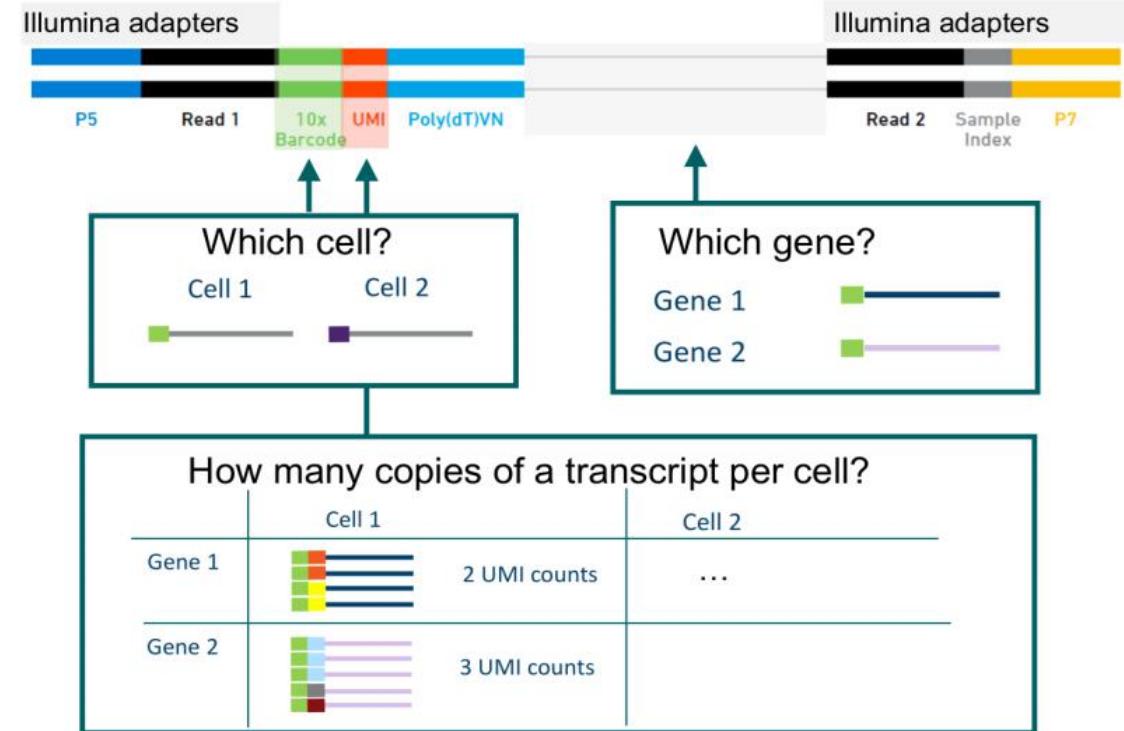
- Input: Single cells in suspension + 10x Gel Beads and Reagents
- Output: Digital gene expression profiles from every partitioned cell

Source: 10x genomics

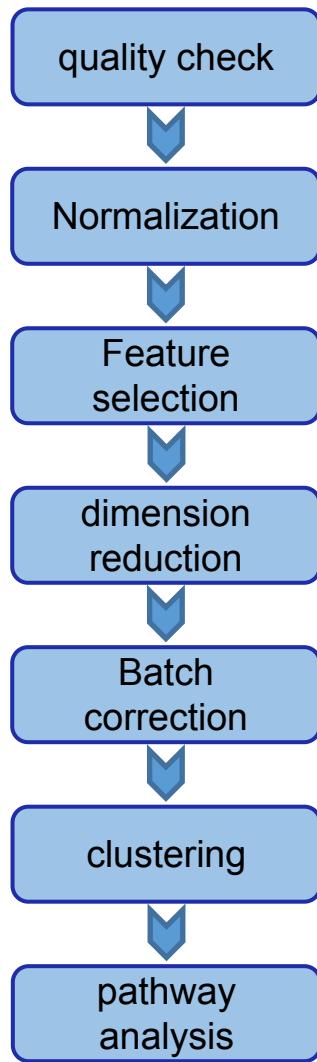
# 10X single cell RNA analysis



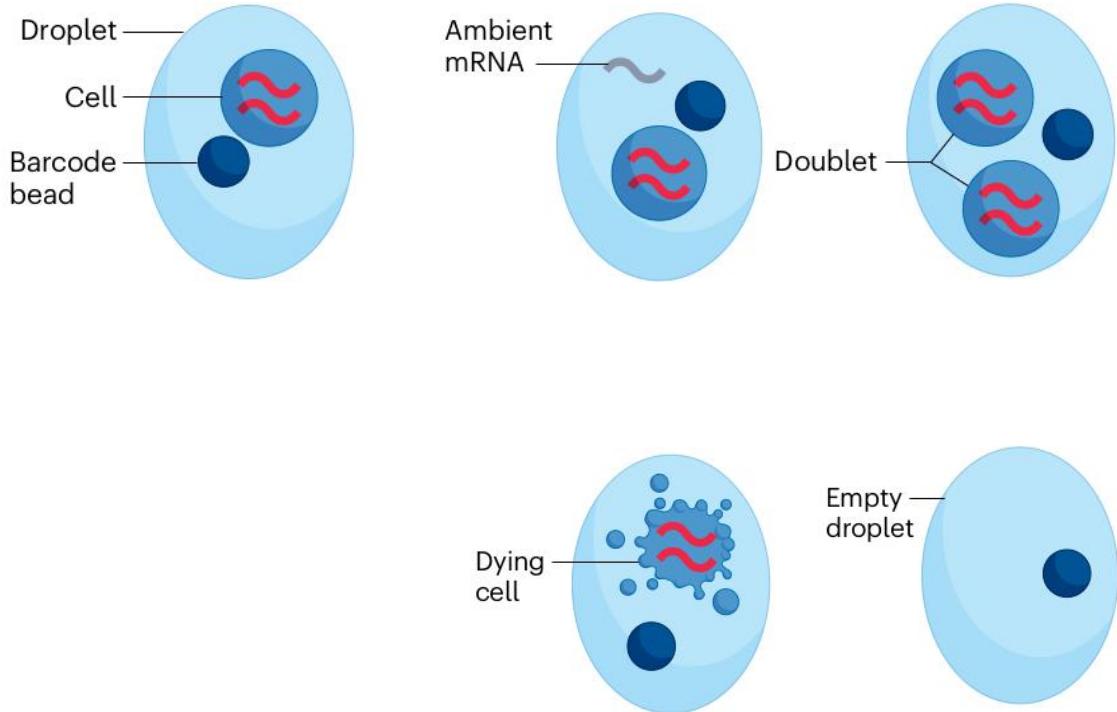
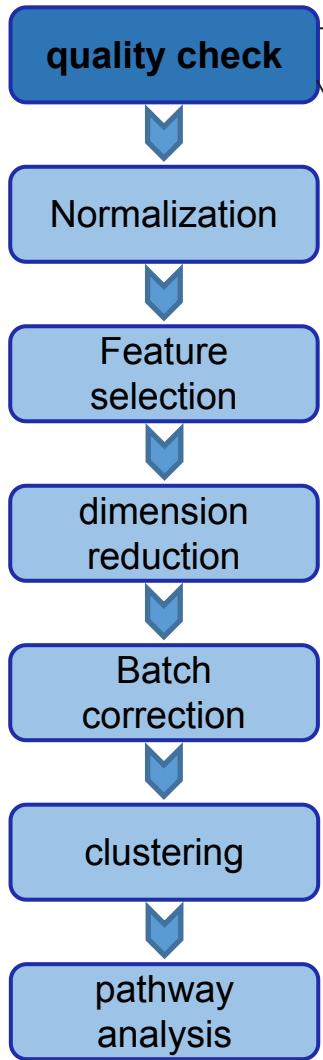
Source: 10x genomics



# Single cell pipeline

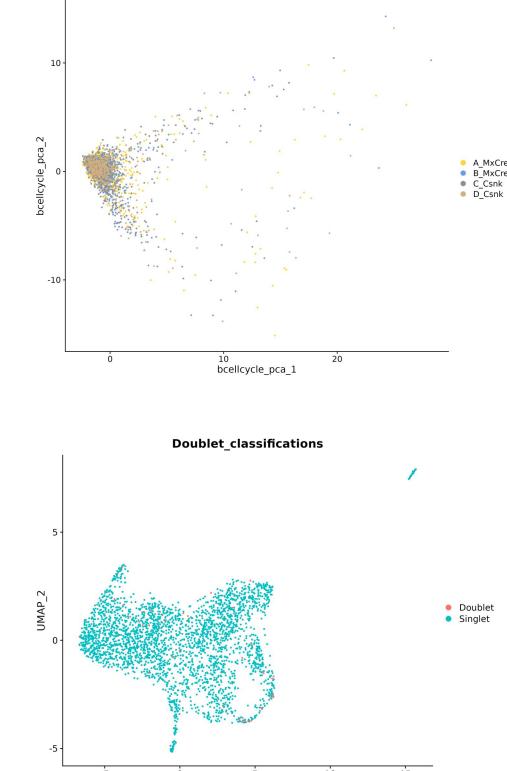
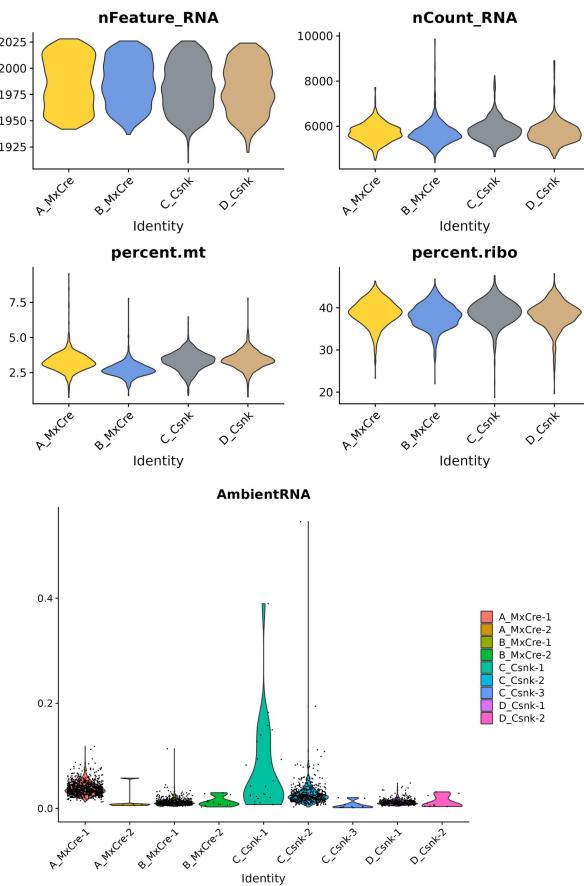
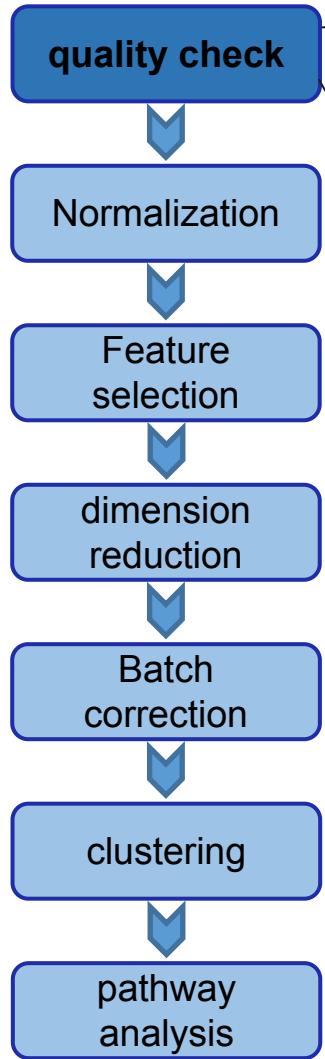


# Single cell pipeline

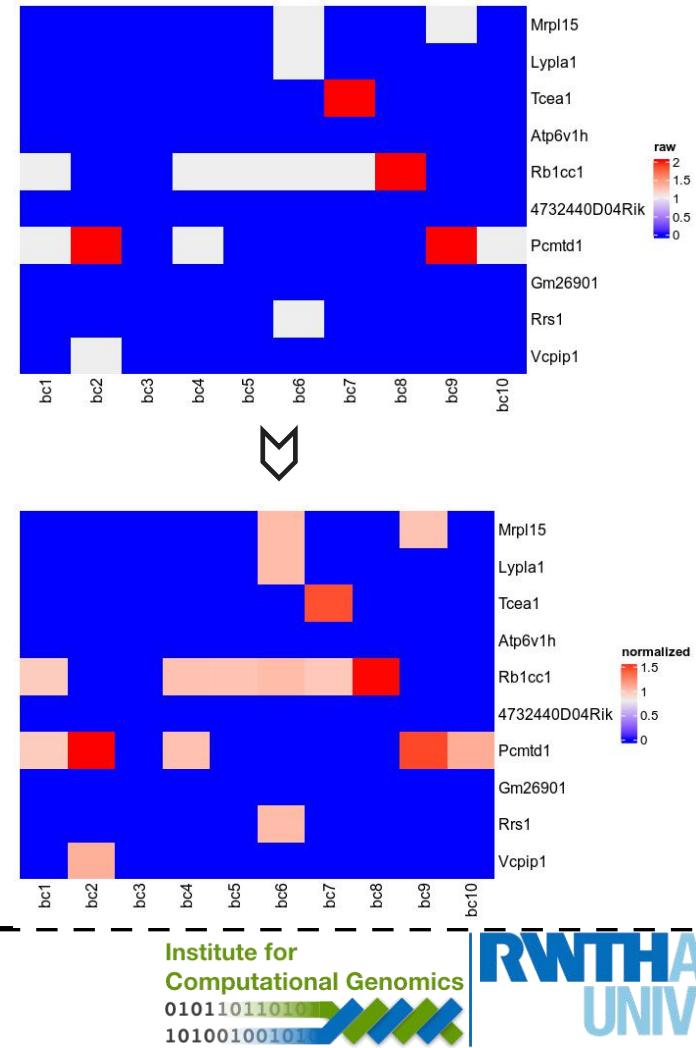
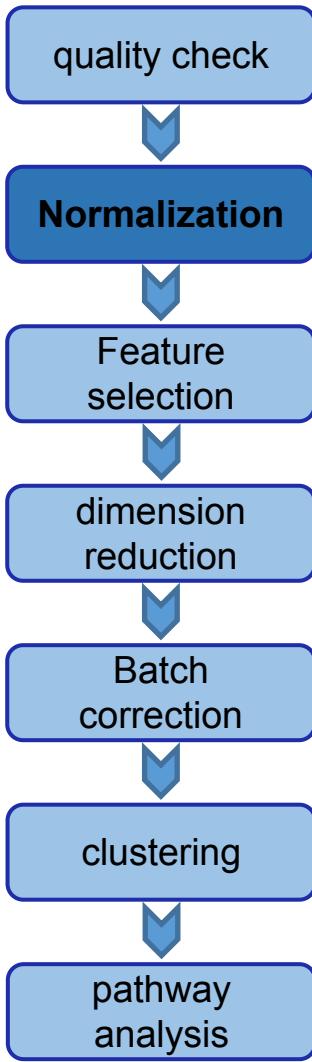


Heumos, Lukas, et al. "Best practices for single-cell analysis across modalities." *Nature Reviews Genetics* (2023): 1-23.

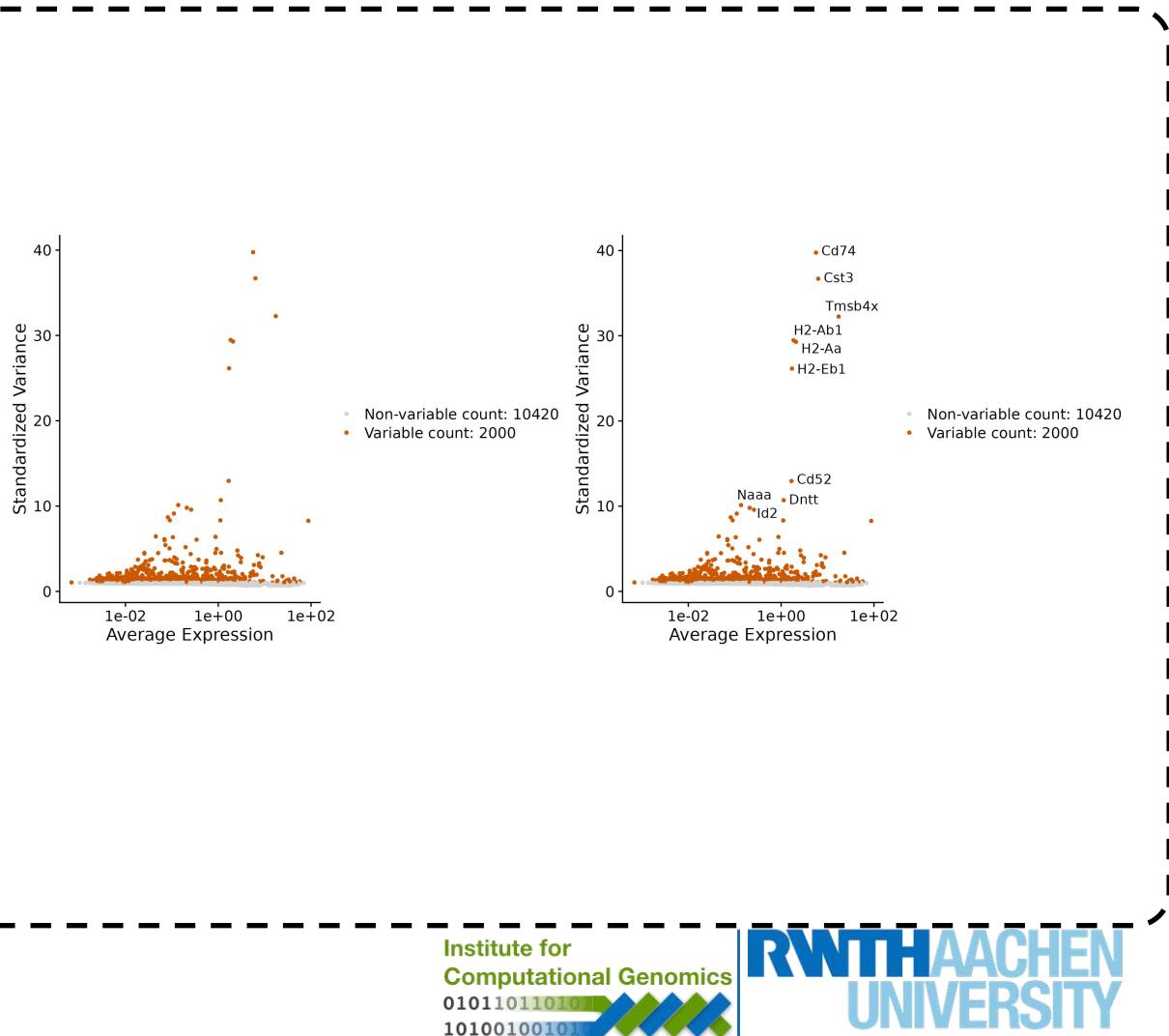
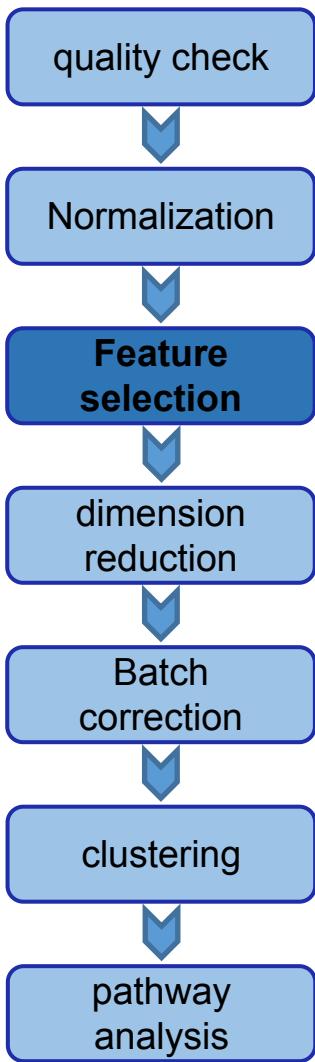
# Single cell pipeline



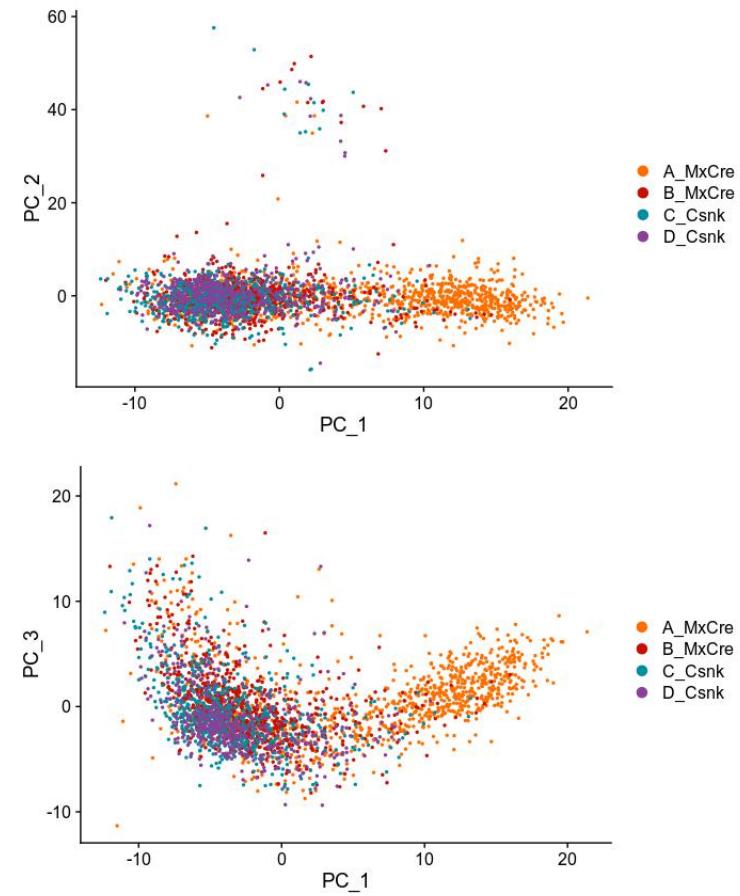
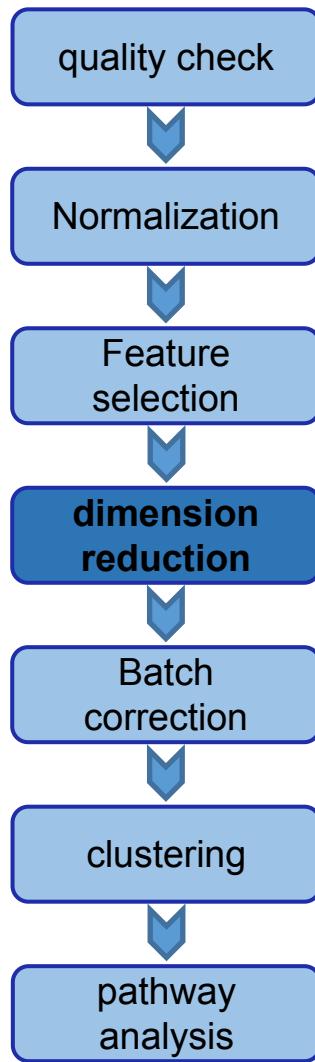
# Single cell pipeline



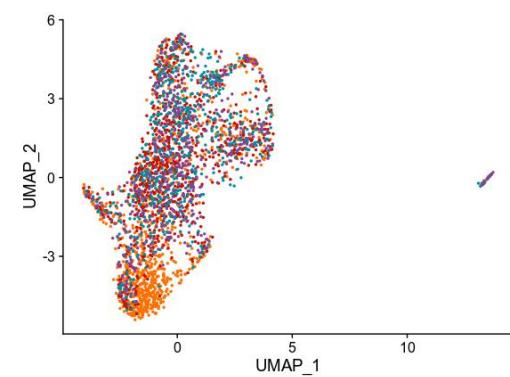
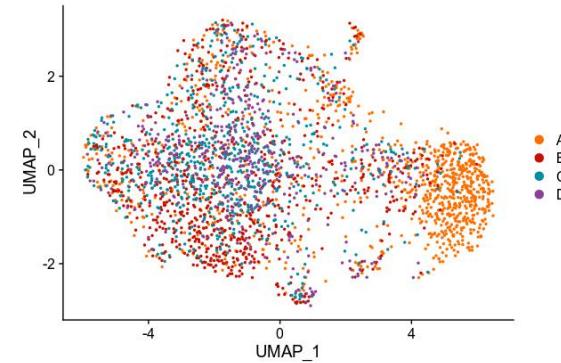
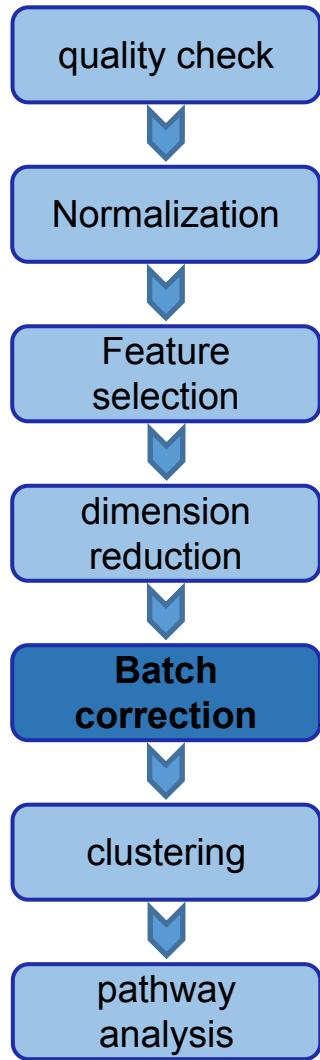
# Single cell pipeline



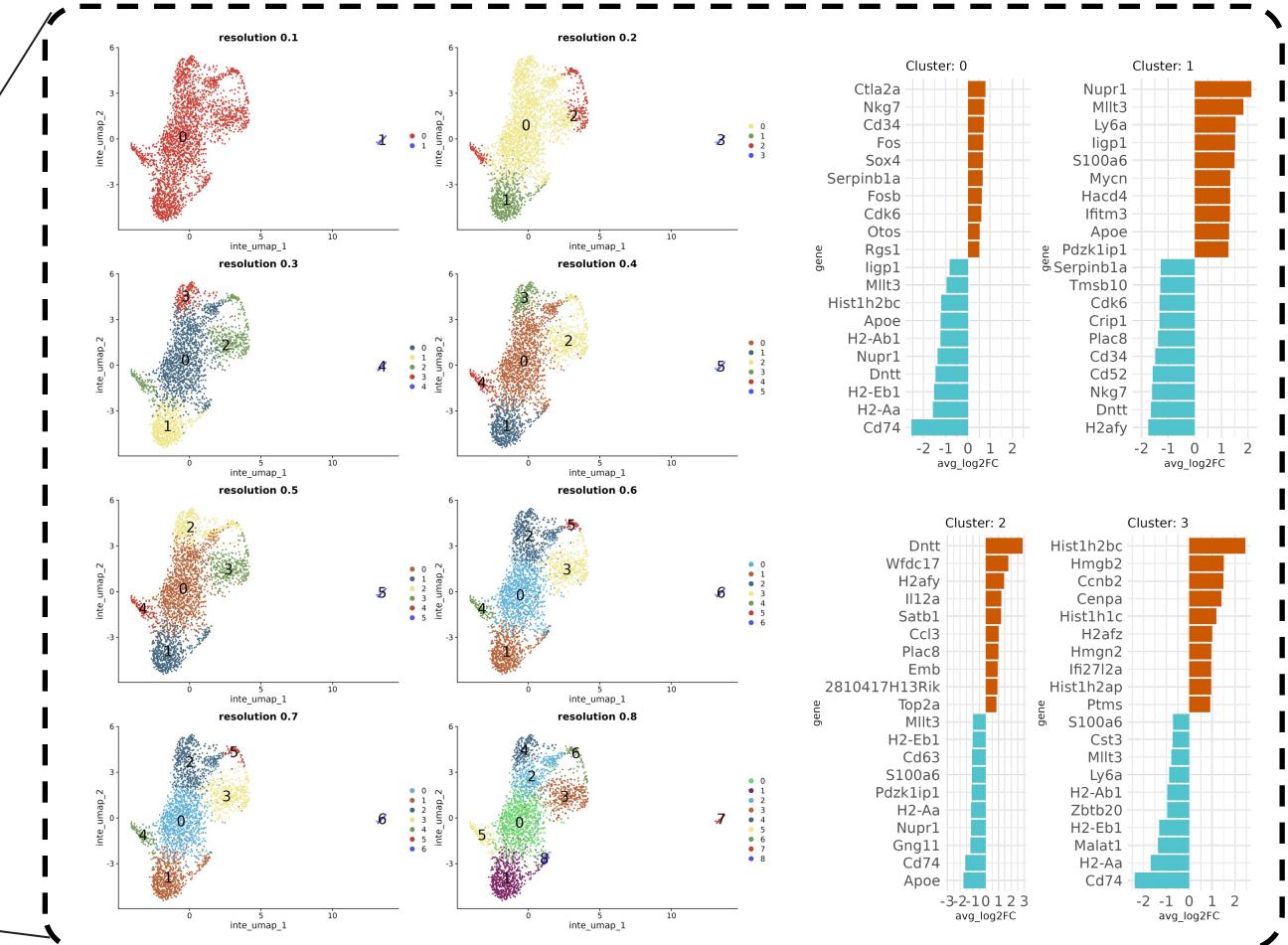
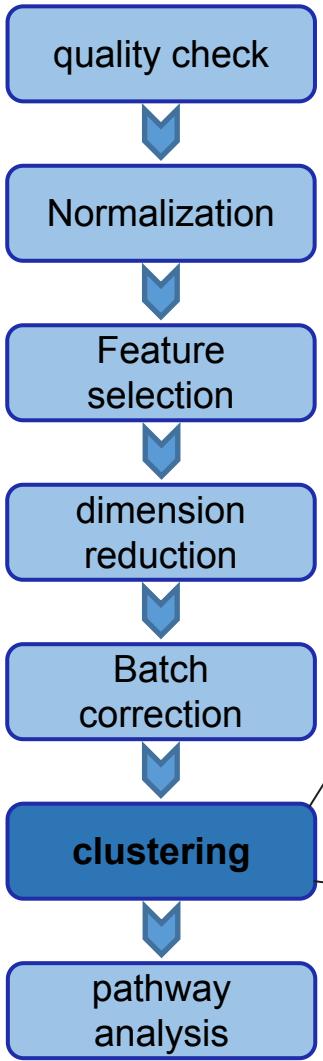
# Single cell pipeline



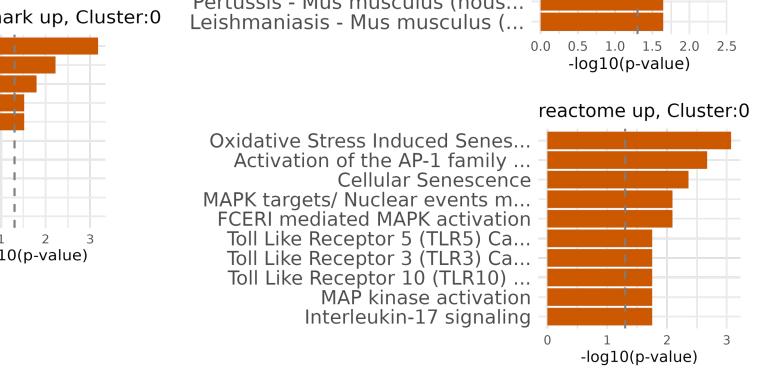
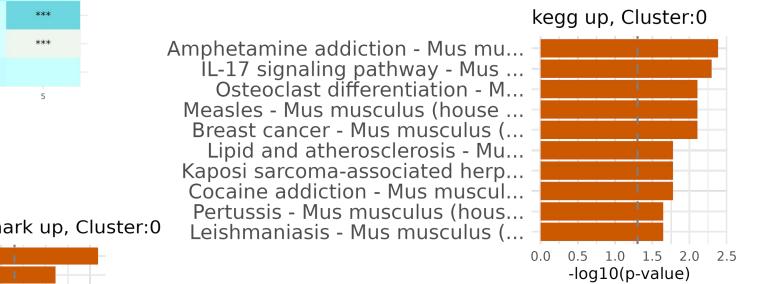
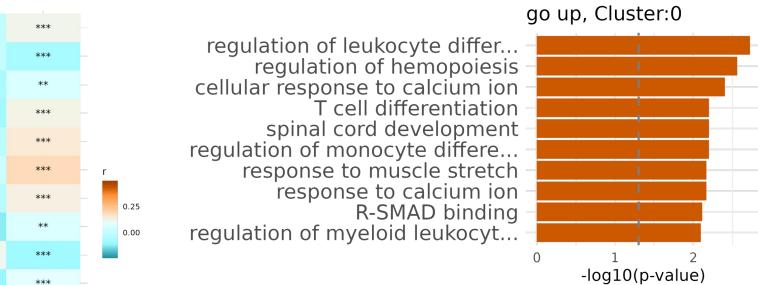
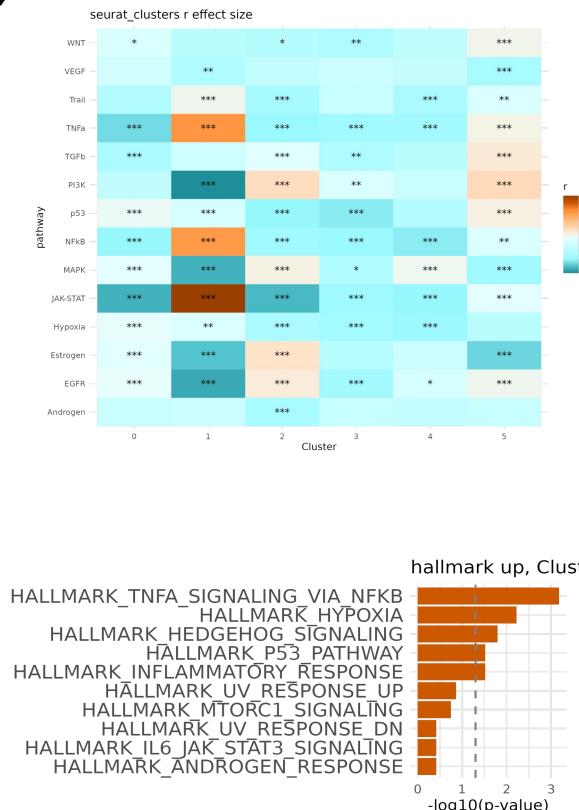
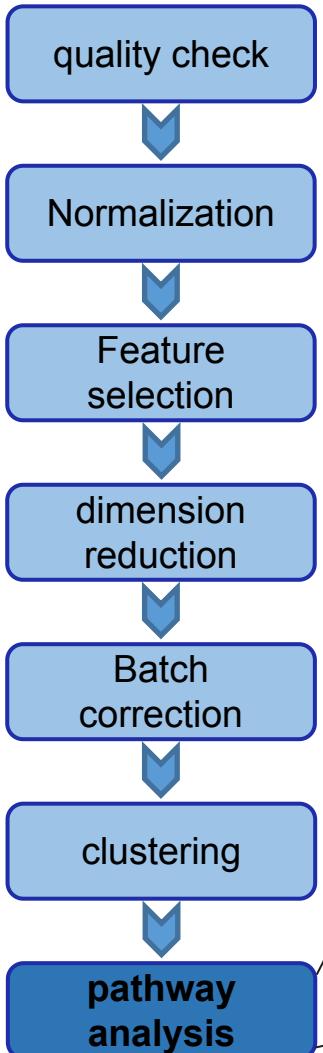
# Single cell pipeline



# Single cell pipeline



# Single cell pipeline

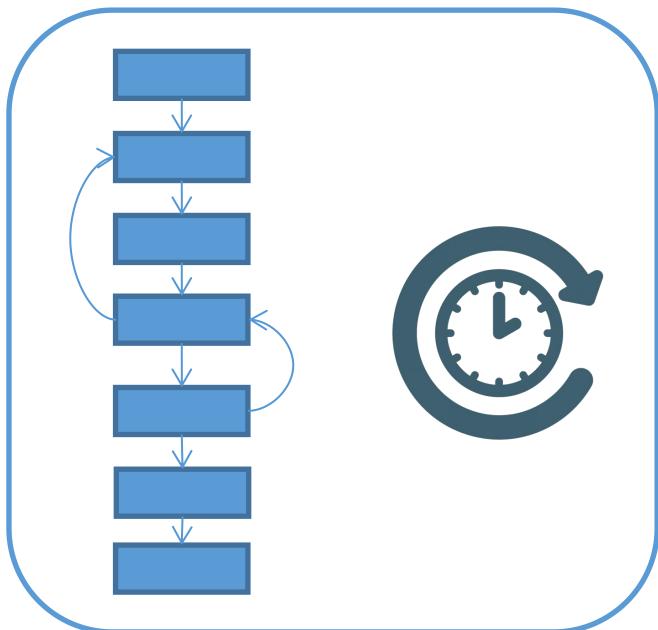


# Outline

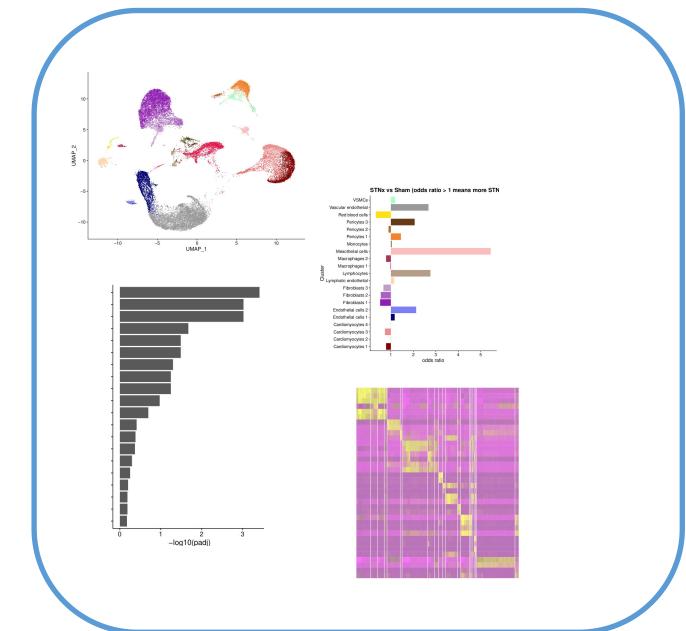
- Overview of data analysis in scRNAseq data
- **Single cell pipeline introduction**
  - overview
  - output example
  - quick start
  - parameter settings
  - frequently asked questions

# Single cell pipeline introduction -- overview

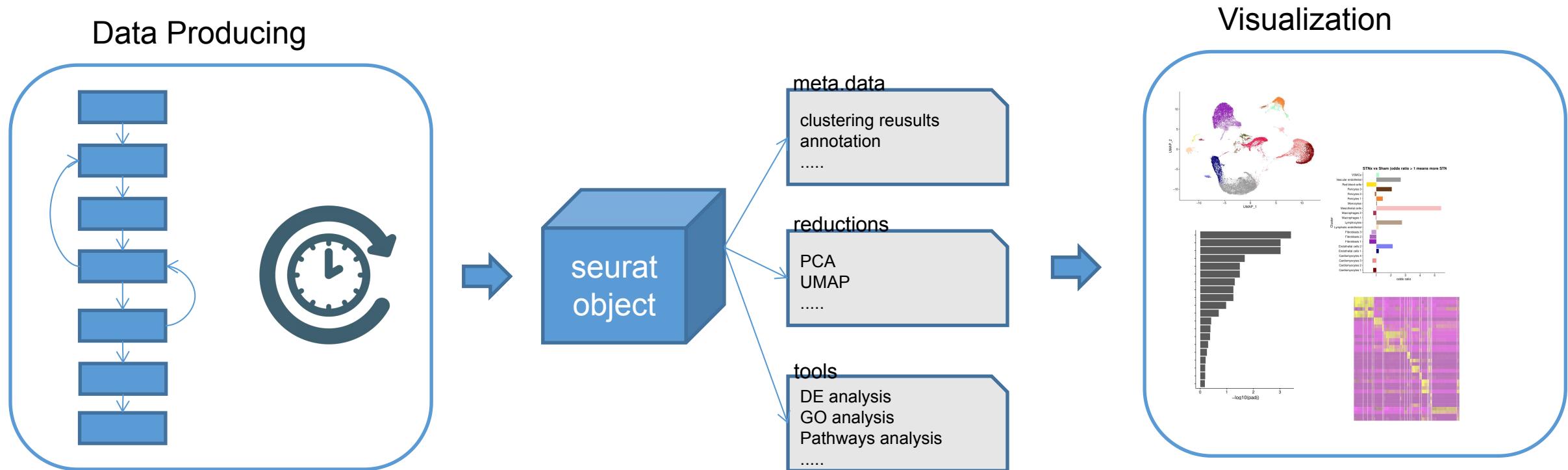
Data Producing



Visualization



# Single cell pipeline introduction -- overview



# Outline

- Overview of data analysis in scRNASeq data
- **Single cell pipeline introduction**
  - overview
  - **output example**
  - quick start
  - parameter settings
  - frequently asked questions

# output example

## Mouse Blood project Analysis Pipeline

### 1. Data Quality Check

- Data quality
- Data quality existing data
- Ambient RNA
- Doublet Detection

### 2. Batch Clustering Results

- seurat\_clusters
- harmony\_clusters
- Marker genes with different resolutions - Integration: seurat

### 3. Final Clustering Results

- clusters - Integration: seurat

### 4. Marker Genes & GO & pathway analysis

- External Markers
- DE & GO
  - DE & GO
    - DE-excel
    - GO-UP-excel
    - GO-DOWN-excel
- Genesets
  - Genesets
- progeny

# output example

## Mouse Blood project Analysis Pipeline

### 1. Data Quality Check

- Data quality
- Data quality existing data
- Ambient RNA
- Doublet Detection

### 2. Batch Clustering Results

- seurat\_clusters
- harmony\_clusters
- Marker genes with different resolutions - Integration: seurat

### 3. Final Clustering Results

- clusters - Integration: seurat

### 4. Marker Genes & GO & pathway analysis

- External Markers
- DE & GO
  - DE & GO
    - DE-excel
    - GO-UP-excel
    - GO-DOWN-excel
- Genesets
  - Genesets
- progeny

- Hallmark
  - hallmark
    - Hallmark-UP-excel
    - Hallmark-DOWN-excel

- KEGG
  - KEGG
    - KEGG-UP-excel
    - KEGG-DOWN-excel

- Reactome
  - Reactome
    - Reactome-UP-excel
    - Reactome-DOWN-excel

- Interactive UMAPs
  - Interactive UMAPs

### 5. Differential Expression & GO analysis(between groups)

- DE&GO pages
  - MxCre.vs.Csnk
- DE FILES
  - DE MxCre.vs.Csnk.xlsx
- GO FILES
  - GO UP MxCre.vs.Csnk.xlsx
  - GO DOWN MxCre.vs.Csnk.xlsx

### 6. Pathway analysis (between groups)

- Genesets
  - Genesets stage

# Outline

- Overview of data analysis in scRNASeq data
- **Single cell pipeline introduction**
  - overview
  - output example
  - **quick start**
  - parameter settings
  - frequently asked questions

# quick start

```
git clone https://github.com/CostaLab/scrna_seurat_pipeline.git
```

```
Rscript packages_install.R
```

```
cd scrna_seurat_pipeline
```

```
cp conf/config.R conf/config_toy.R      ## toy configuration file
```

```
cp run_example.sh run_toy.sh           ## Please edit run_toy.sh to fit your environment
```

```
sh run_toy.sh toy
```

```
cp run_viz_example.sh run_viz_toy.sh    ## Please edit run_viz_toy.sh to fit your environment
```

```
sh run_viz_toy.sh toy
```

# quick start

```
git clone https://github.com/CostaLab/scrna_seurat_pipeline.git
```

```
Rscript packages_install.R
```

```
cd scrna_seurat_pipeline
```

```
cp conf/config.R conf/config_toy.R ## toy configuration file
```

```
cp run_example.sh run_toy.sh ## Please edit run_toy.sh to fit your environment
```

```
sh run_toy.sh toy
```

```
cp run_viz_example.sh run_viz_toy.sh ## Please edit run_viz_toy.sh to fit your environment
```

```
sh run_viz_toy.sh toy
```

# quick start

- Mainly split **execution plan** to 3 main step
  - scrna\_phase\_preprocess
  - scrna\_phase\_clustering
  - scrna\_phase\_comparing

# quick start

- Mainly split **execution plan** to 3 main step
  - **scrna\_phase\_preprocess**
    - Ambient RNA detection
    - doublets detection
    - **Filteration**
    - etc.
  - **scrna\_phase\_clustering**
  - **scrna\_phase\_comparing**

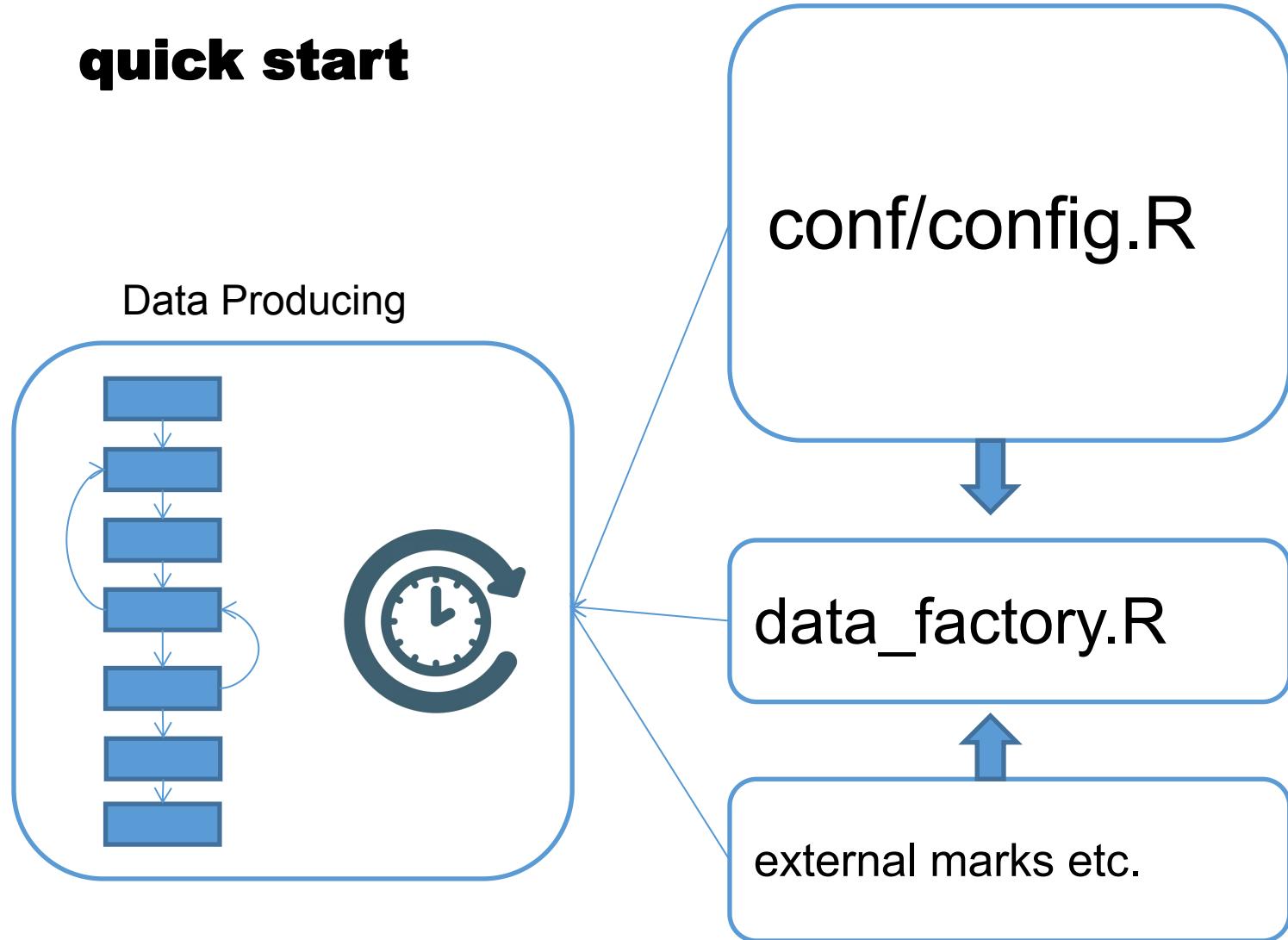
# quick start

- Mainly split **execution plan** to 3 main step
  - scrna\_phase\_preprocess
  - **scrna\_phase\_clustering**
    - Batch correction
    - Clustering
    - MCA, HCL annotation
    - etc.
  - scrna\_phase\_comparing

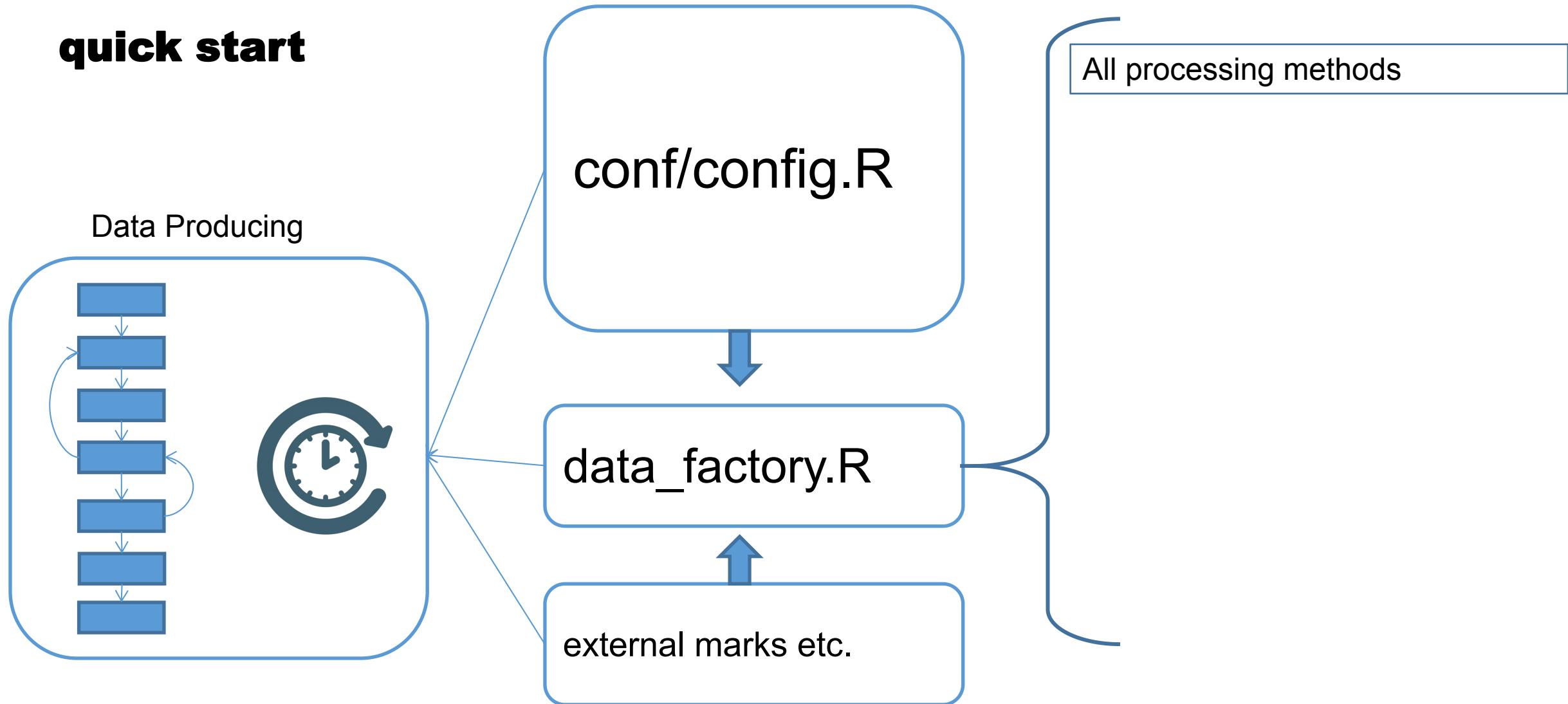
# quick start

- Mainly split **execution plan** to 3 main step
  - scrna\_phase\_preprocess
  - scrna\_phase\_clustering
  - **scrna\_phase\_comparing**
    - fishertest etc.
    - progeny analysis
    - DE analysis
    - GO, KEGG, Reactome, HALLMARK
    - etc.

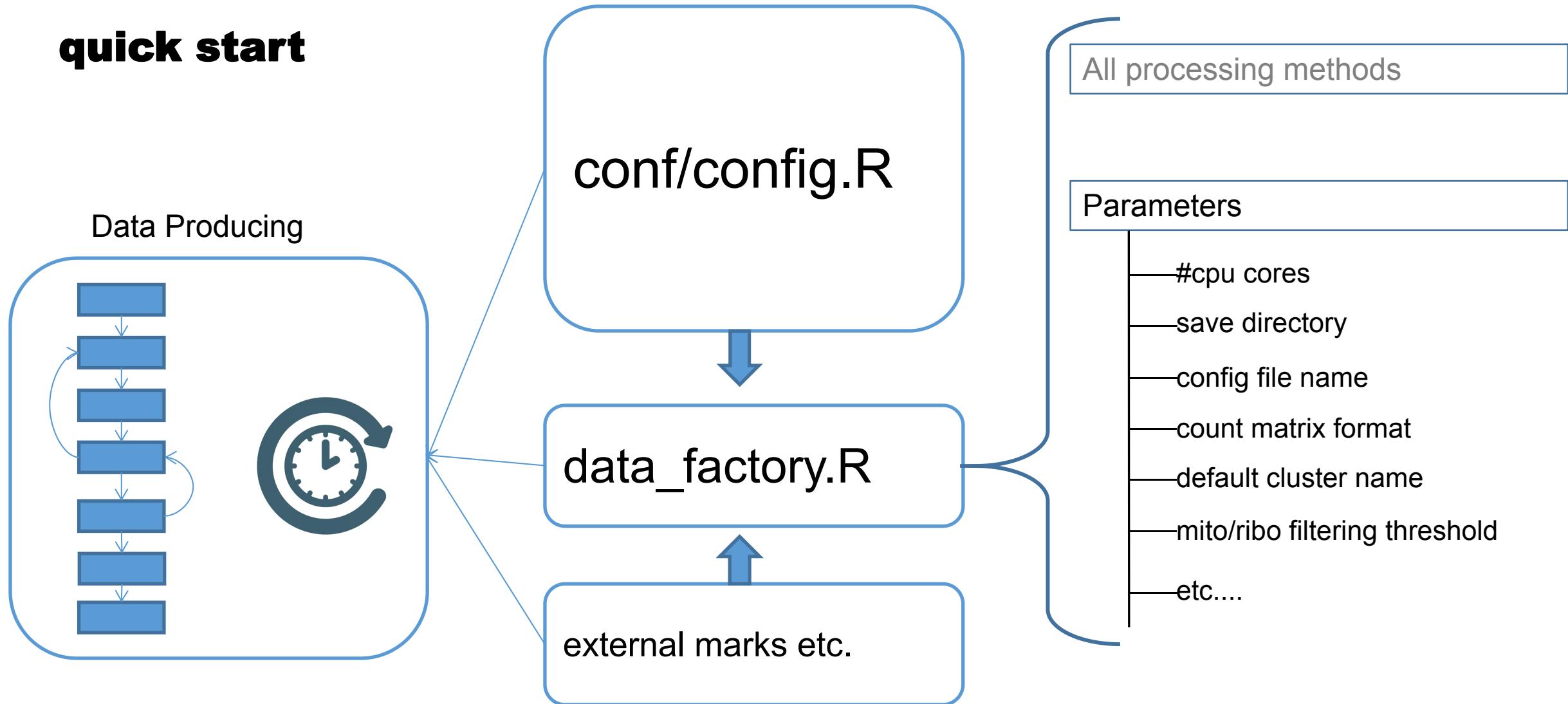
# quick start



# quick start

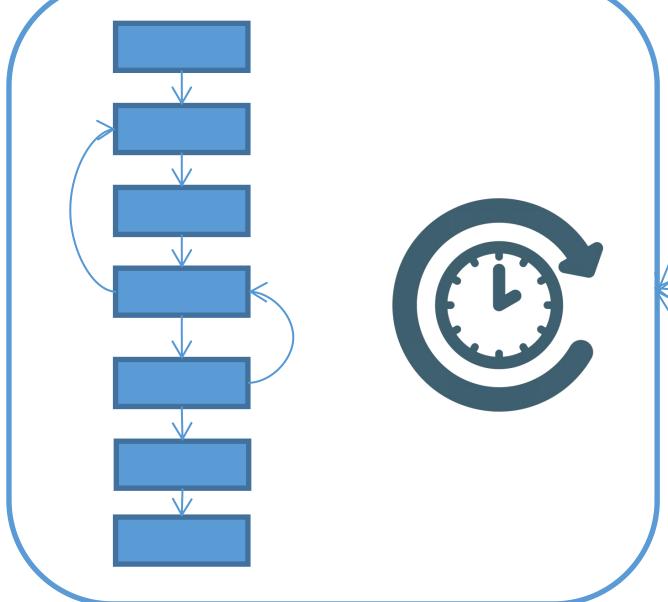


# quick start



# quick start

Data Producing



# run\_toy.sh

All processing methods

Parameters

- #cpu cores
- save directory
- config file name
- count matrix format
- default cluster name
- mito/ribo filtering threshold
- etc....

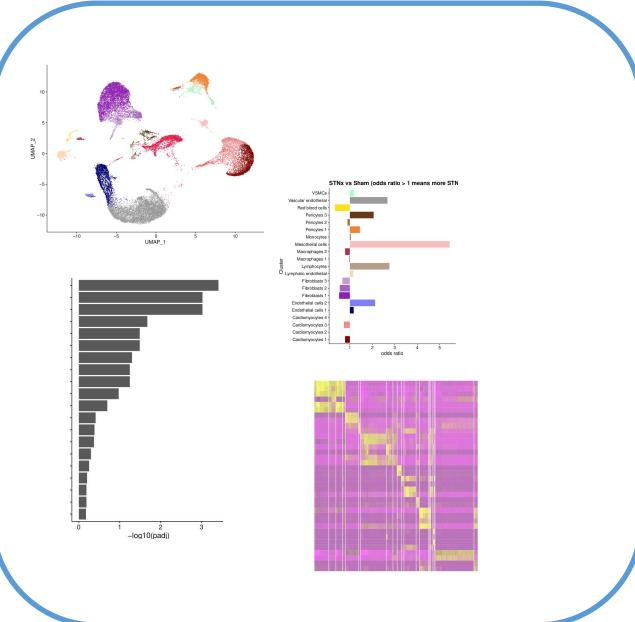
# quick start

- conf/config.R
- excels generated
- Seurat Objects

run\_viz\_toy.sh

- 1\_quality\_report.Rmd
- 2\_clustering.Rmd
- 2\_clusters\_DEs.Rmd
- 3\_DE\_GO-analysis.Rmd
- 3\_external\_markers.Rmd
- make\_report.R
- singleton\_clustering.Rmd
- template/
  - DE-GO-vs.template
  - Genesets-vs.template
  - index.template

## Visualization



report

# Outline

- Overview of data analysis in scRNAseq data
- **Single cell pipeline introduction**
  - overview
  - output example
  - quick start
  - **parameter settings**
  - frequently asked questions

# Outline

- Overview of data analysis in scRNAsq
- Single cell pipeline introduction
  - overview
  - output example
  - quick start
  - parameter settings
  - frequently asked questions

```
vim config.R 99x58
### -----Initial info-----
PROJECT = "Mouse Blood project" ## set project name
ORGAN = 'Blood'                 #For external annotation. Options: see below(External Organs)
SPECIES = "Mouse"                #For external annotation. Options: Human, Mouse
MCA_NAME = "Bone-Marrow" #For MCA annotation.      Options: check http://bis.zju.edu.cn/M
HCL_NAME = "Adult-Bone-Marrow-CD34P" #For HCL annotation.

# filtering params when create seurat object
MINCELLS = 5
MINGENES = 50

INTEGRATION_OPTION = "seurat" ### or harmony

### ----- Data SRC -----
ANNOTATION_EXTERNAL_FILE = "external/Human_and_mouse_cell_markers-Markers.tsv"

## If genesets you need are not included, please attach your geneset to the gmt.gz file.
MSigDB_GENESET_HUMAN_GMT_FILE = "external/Human_msigdb.v7.2.symbols.gmt.gz"

data_src = c(
  A_MxCre     = "data/A_MxCre",
  B_MxCre     = "data/B_MxCre",
  C_Csnk      = "data/C_Csnk",
  D_Csnk      = "data/D_Csnk"
)

##----- SET REPLICATE GROUP -----
stage_lst = c(
  A_MxCre     = "MxCre",
  B_MxCre     = "MxCre",
  C_Csnk      = "Csnk",
  D_Csnk      = "Csnk"
)
```

# Outline

- Overview of data analysis in scRNAseq data

- **Single cell pipeline introduction**

- overview
- output example
- quick start
- **parameter settings**
- frequently asked questions

```
## 0. omit,      1. calc & save,      2. load
conf = c(
  scrna_phase_preprocess      = 1, ## quali
  scrna_phase_clustering      = 1, ## integ
  scrna_phase_comparing       = 1, ## DE GO
  scrna_cluster_annotation    = 0, ## Annot
  scrna_clusterwise_xcell     = 2, ## remov
  scrna_del_mitogenes         = 0, ## !!!DA
  scrna_merge_clusters        = 0, ## merge
  scrna_remove_clusters       = 0, ## remov
  scrna_remove_recluster      = 0) ## remov
```

# Outline

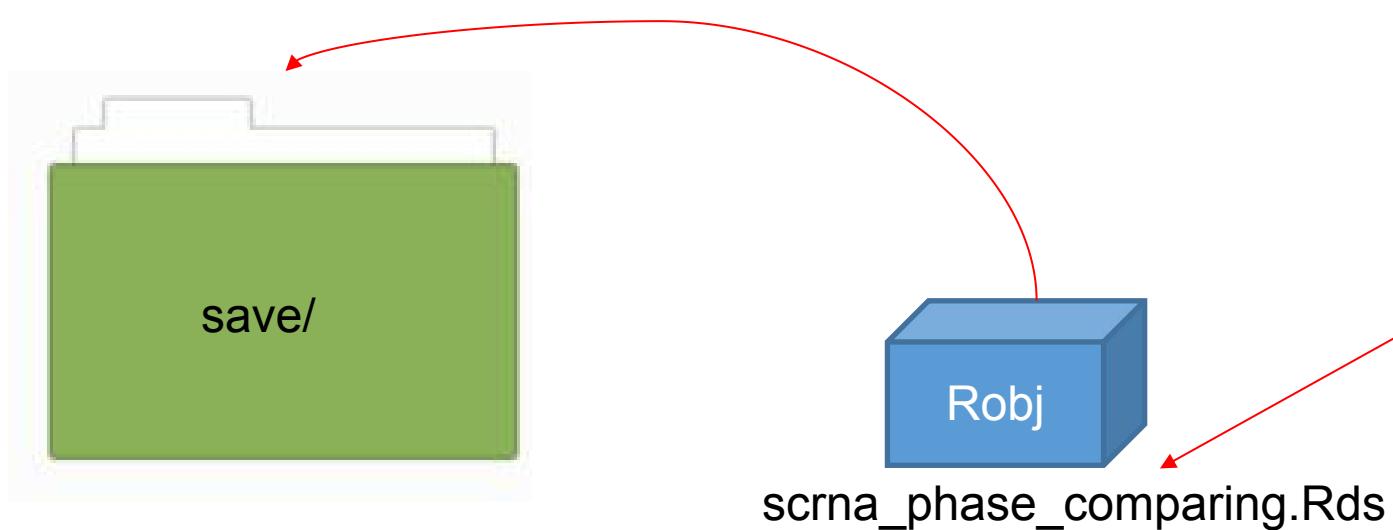
- Overview of data analysis in scRNASeq data
- **Single cell pipeline introduction**
  - overview
  - output example
  - quick start
  - parameter settings
  - **frequently asked questions**

## frequently asked questions

- Is it possible to do further analysis given an existed Seurat Object?

# frequently asked questions

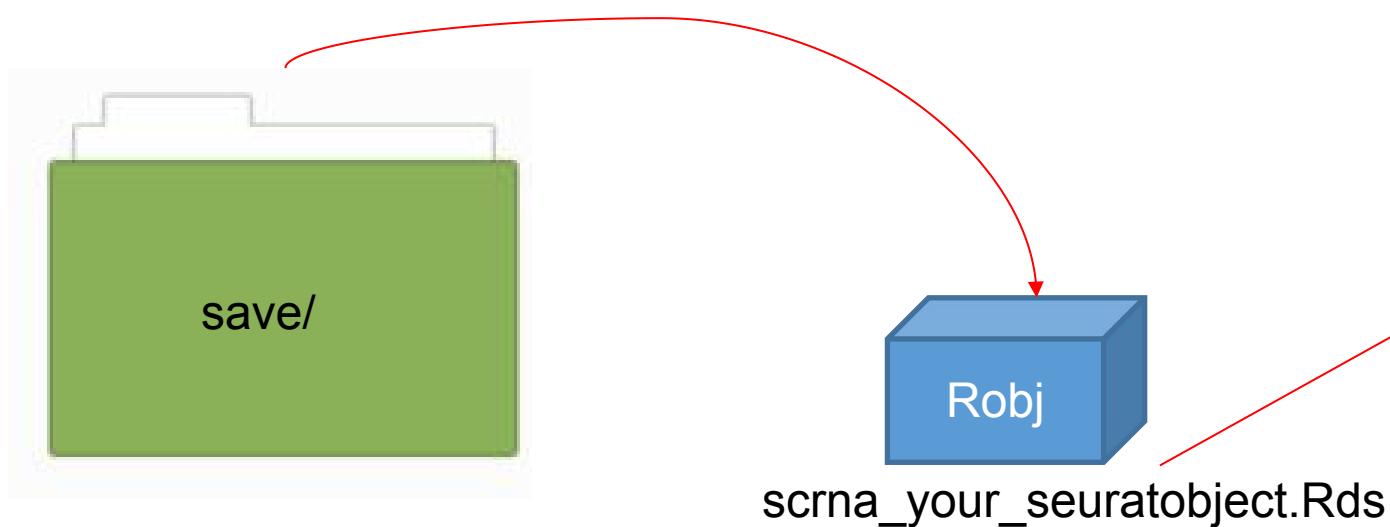
- Is it possible to do further analysis given an existed Seurat Object?



```
## 0. omit,      1. calc & save,      2. load
conf = c(
  scrna_phase_preprocess = 1, ## quali
  scrna_phase_clustering = 1, ## integ
  scrna_phase_comparing = 1, ## DE GO
  scrna_cluster_annotation = 0, ## Annot
  scrna_clusterwise_xcell = 2, ## remov
  scrna_del_mitogenes = 0, ## !!!DA
  scrna_merge_clusters = 0, ## merge
  scrna_remove_clusters = 0, ## remov
  scrna_remove_recluster = 0) ## remov
```

# frequently asked questions

- Is it possible to do further analysis given an existed Seurat Object?



```
## 0. omit,      1. calc & save,      2. load
conf = c(
  scrna_phase_preprocess = 0, ## qua
  scrna_phase_clustering = 0, ## ini
  scrna_your_seuratobject = 2, ## DE
  scrna_phase_clustering = 1, ## ini
  scrna_phase_comparing = 1, ## DE
  scrna_cluster_annotation = 0, ## An
  scrna_clusterwise_xcell = 0, ## ren
  scrna_del_mitogenes = 0, ## !!
  scrna_merge_clusters = 0, ## me
  scrna_remove_clusters = 0, ## ren
  scrna_remove_recluster = 0) ## ren
```

## frequently asked questions

- Is it possible to do further analysis given an Seurat Object?
- Why I failed to load the RObject in the save folder?

# frequently asked questions

- Is it possible to do further analysis given an Seurat Object?
- Why I failed to load the RObject in the save folder?
  - ```
source("R/save_load_helper.R")
scrna <- load_object("scrna_phase_comparing.Rds")
```

## frequently asked questions

- Is it possible to do further analysis given an Seurat Object?
- Why I failed to load the RObject in the save folder?
- Why not using nextflow or snakemake?

# frequently asked questions

- Is it possible to do further analysis given an Seurat Object?
- Why I failed to load the RObject in the save folder?
- Why not using nextflow or snakemake?
  - Too many I/O tasks
  - More flexible to do the analysis
  - Nextflow produces too many files leading the RWTH-HPC overloaded

# frequently asked questions

- Is it possible to do further analysis given an Seurat Object?
- Why I failed to load the RObject in the save folder?
- Why not using nextflow or snakemake?
  - Too many I/O tasks
  - More flexible to do the analysis
  - Nextflow produces too many files leading the RWTH-HPC over

```
vim phase.ini 78x48
1 [phase_preprocess] = 1
2 scrna_rawdata = 1
3 scrna_ambient_rna = 1
4 scrna_filter = 1
5 scrna_doublet_proportions = 1
6 scrna_preprocess = 1
7 scrna_cellcycle = 1
8 scrna_cycleRegressOut = 1
9 scrna_regressOut = 1
10
11
12
13 [phase_clustering] = 1
14 scrna_integration_harmony = 1
15 scrna_integration_seurat = 1
16 scrna_batchclustering = 1
17 scrna_batch_markergenes = 1
18 scrna_clustering = 1
19 scrna_fishertest_inte_clusters = 1
20 scrna_fishertest_clusters = 1
21 scrna_proptest_clusters = 1
22 scrna_MCAannotate = 1
23 scrna_ExternalAnnotation = 1
24 scrna_HCLannotate = 1
25 scrna_MAGIC = 1
```

## frequently asked questions

- Is it possible to do further analysis given an Seurat Object?
- Why I failed to load the RObject in the save folder?
- Why not using nextflow or snakemake?
- The pipeline require too much memory

# frequently asked questions

- Is it possible to do further analysis given an Seurat Object?
- Why I failed to load the RObject in the save folder?
- Why not using nextflow or snakemake?
- The pipeline require too much memory
  - Please set `--allinone=FALSE` in `run_toy.sh`

# **useful links**

- **source code**
  - [https://github.com/CostaLab/scrna\\_seurat\\_pipeline](https://github.com/CostaLab/scrna_seurat_pipeline)
- **tutorial**
  - [https://costalab.github.io/scrna\\_seurat\\_pipeline/test/tutorial.html#run-data-producing](https://costalab.github.io/scrna_seurat_pipeline/test/tutorial.html#run-data-producing)
- **Link**
  - [https://github.com/CostaLab/scrna\\_seurat\\_pipeline/blob/master/docs/A\\_comprehensive\\_singlecell\\_RNA\\_analysis\\_pipeline.pdf](https://github.com/CostaLab/scrna_seurat_pipeline/blob/master/docs/A_comprehensive_singlecell_RNA_analysis_pipeline.pdf)

# Thanks

## Q&A