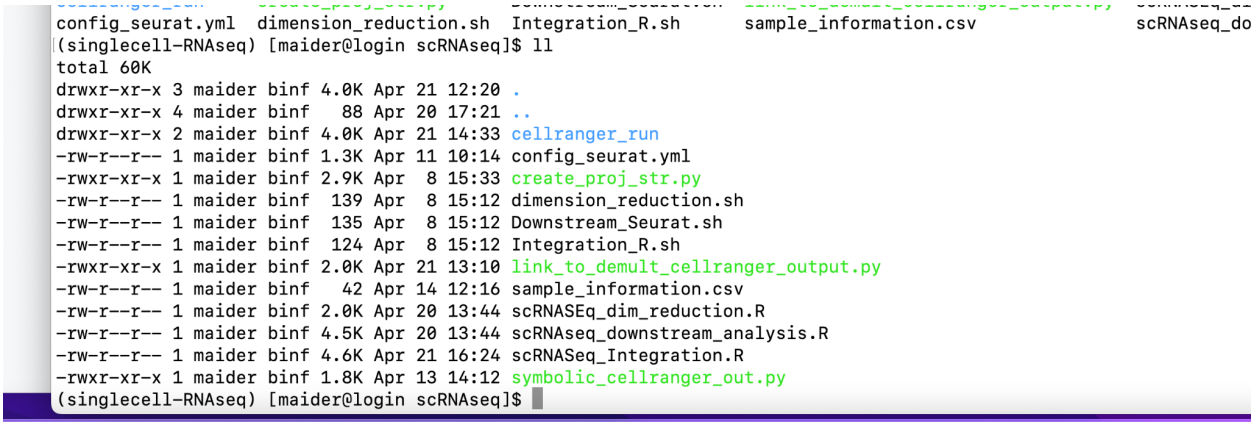
singlecellRNA-seq analysis

Pipeline consists of two inputs: metadata file with sample names and condition
output from cellranger: barcodes.tsv.gz features.tsv.gz matrix.mtx.gz.

Three R scripts for Integration, dimension reduction and downstream analysis and sh scripts to launch these three files.

One config file

Two python scripts: to create needed output folders
to create nedded folder configuration older with output files from cell ranger. Depending on which kind of the data we received: multiplexed or demultiplexed this Python script will be different.



If you are running in the server first:
Ssh maider@ciilogin.c2b2.columbia.edu

And then:

qlogin -l mem=30G

cd /share/data/scRNAseq_test

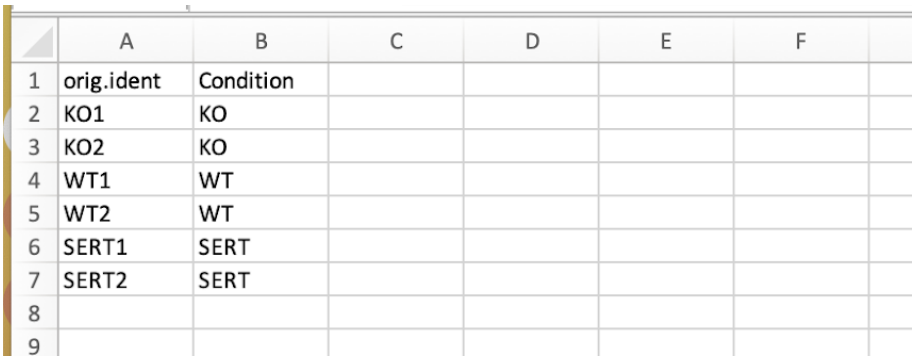
cp ~/pipelines/scRNAseq/*

The starting working directory will look as follows:

```
[maider@login scRNAseq]$ cp /share/data/scRNAseq_test/*.sh .
[maider@login scRNAseq]$ ls
cellranger_run      create_proj_str.py      Downstream_Seurat.sh  scRNASEq_dim_reduction.R        scRNASeq_Integration.R
config_seurat.yml  dimension_reduction.sh  Integration_R.sh      scRNAseq_downstream_analysis.R  symbolic_cellranger_out.py
[maider@login scRNAseq]$
```

The only input needed to upload at this point from the computer is the csv with the sample metadata.

Sample metadata information will be in csv format as follows. Headers of sample name and condition have to be the same. More columns can be added. Sample name has to be the same as the output folder names in cellranger.



	A	B	C	D	E	F
1	orig.ident	Condition				
2	KO1	KO				
3	KO2	KO				
4	WT1	WT				
5	WT2	WT				
6	SERT1	SERT				
7	SERT2	SERT				
8						
9						

scp sample_informatio.csv maider@ciilogin.c2b2.columbia.edu:/path_to_working_directory/

Config.yml has the following variables:

```yaml
default:
  data_location: "data"
  cellranger_out: "/share/data/RNA_Seq/10X/Raw_fastq/" #output location for all samples
  cellranger_option: "COUNT" #can be MULTI or COUNT
  metadata_file: "seurat_sample_metadata.csv"
  condition_reference: "HC" #condition to use as reference
  quality_location: "Quality_figures/" #output of quality figures
  seurat_object: "Seurat_objects/" #output of Seurat object
  seurat_object_name: "Seurat_object.RData" #name of the output Seurat object
  seurat_object_dim: "dim_Seurat_object.RData" #name of the output Seurat obejct with dimensio reduction
  integration_folder: "Integration_results/" #outout of downstream integration figures and tables
  DE_folder: "DE_Analysis/" #output of DE analysis
  enrichment_folder: "Enrichment_files/" #output of files to load in GSEA

#These all are for filtering steps
  min_cells: 3
  min_features: 200
  nFeature_max: 8000
  nFeature_min: 50
  nCount_max: 60000
  nCount_min: 20
  percent_mit: 10
  HVF_selection: 2000

#Number of dimensions to use for reduction
  npcs: 30
  resolution: 0.3
  Cell_type_markers: D3D,KLRF1,FCGR3A,CD14,CD1C,IL3RA,IGKC,MYL9,HBB,CD79A

  markers_outfile: "Markers.csv" #output name of markers

#Downstream analysis
  min.pct: 0.1
  log.fc_threshold_markers: 0.25
  log.fc_threshold_DE: 0.25

```

Then we run the python script to create the directories. In the server you will run:
python create_proj_str.py -p /path

```
(singlecell-RNAseq) [maider@login scRNAseq]$
(singlecell-RNAseq) [maider@login scRNAseq]$ python create_proj_str.py --help
usage:  This script checks if a project directory and sub-directory exists,
        and creates them if necessary. Please provide a project location. Exiting....

Process command line arguments.

optional arguments:
  -h, --help            show this help message and exit
  -p PATH, --path PATH  project path
```

```
(singlecell-RNAseq) [maider@login MECFS_test]$ python3 create_proj_str.py -p .
Project Location: .

Checking/Creating sub-directory: ./Quality_figures
Finished creating the./Quality_figures

Checking/Creating sub-directory: ./Seurat_objects
Finished creating the./Seurat_objects

Checking/Creating sub-directory: ./Integration_results
Finished creating the./Integration_results

Checking/Creating sub-directory: ./DE_Analysis
Finished creating the./DE_Analysis

Checking/Creating sub-directory: ./Enrichment_files
Finished creating the./Enrichment_files
```

Then we will copy the output from cellranger. For this we have two python scripts. In the case we received fastq files demultiplexed and we run cellranger count command (go to Commands/Software document) we will use: symbolic_cellranger_out.py.

```
(singlecell-RNAseq) [maider@login scRNAseq]$
(singlecell-RNAseq) [maider@login scRNAseq]$ python symbolic_cellranger_out.py --help
usage:  This script create the directory structure need to run Seurat R script. Exiting....

Process command line arguments.

optional arguments:
  -h, --help            show this help message and exit
  -cp CELLRANGER_PATH, --cellranger_path CELLRANGER_PATH
                        cellranger output path
  -m FILE, --metadata FILE
                        metadata file
```

If we receive data multiplexed and we had to run cellranger multi we will use: link_to_demult_cellranger_output.py

```
(singlecell-RNAseq) [maider@login scRNAseq]$ python link_to_demult_cellranger_output.py --help
usage:  This script create the directory structure need to run Seurat R script. Exiting....

Process command line arguments.

optional arguments:
  -h, --help            show this help message and exit
  -cp CELLRANGER_PATH, --cellranger_path CELLRANGER_PATH
                        cellranger output path
  -m FILE, --metadata FILE
                        metadata file
```

In both cases:

python symbolic_cellranger_out.py -cp '/share/data/RNA_Seq/10X/Raw_fastq/pbmc_1k_v3_fastqs_2/' -m sample_information.csv

The data folder format will be as follows:

```
(base) mastorkia@maiders-mbp singlecell_MECFS % cd data
(base) mastorkia@maiders-mbp data % ls
KO1     KO2     SERT1   SERT2   WT1     WT2
(base) mastorkia@maiders-mbp data % cd KO1
(base) mastorkia@maiders-mbp KO1 % ls
barcodes.tsv.gz features.tsv.gz matrix.mtx.gz
(base) mastorkia@maiders-mbp KO1 %
```

Once we have all this ready we can launch our scripts.

Go to Seurat_pipeline_documentation