**Background materials:**

* Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data

(<https://www.nature.com/articles/s41596-020-00409-w>)

<https://www.singlecellcourse.org>

* Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics

(<https://www.nature.com/articles/s41576-021-00370-8>)

* Deciphering cell–cell interactions and communication from gene expression

(<https://www.nature.com/articles/s41576-020-00292-x>)

* Neurobiological perspective of 22q11.2 deletion syndrome

(<https://pubmed.ncbi.nlm.nih.gov/31395526/>)

* A breaking news on stem cell modeling

https://www.cam.ac.uk/stories/model-embryo-from-stem-cells

**Snippets of sample data:**

* Data format for R:

<https://github.com/satijalab/seurat/wiki/Seurat>

* Data format for python:

<https://anndata.readthedocs.io/en/latest/>

**Students' desirable skills/background:**

* Programming skills with R (Be familiar with Bioconductor packages) and Python. Experience with software development will be a plus.
* Basic knowledge in biostatistics
* Background knowledge in neuroscience/biology will be a plus

**Clarifications on the goal, methods, deliverables**:

**Project Title:** Analysis disease mechanisms of 22q11 microdeletion syndrome, a severe neuropsychiatric disorder, from single cell RNA sequencing (scRNAseq) data

**Project Goal:**

1. Modularize and standardize a basic scRNAseq analysis pipeline
2. chart the changes in spatial dynamics of 22q11DS

**Problems need your inputs:**

1. **Data organization**
   1. Raw
      1. Data from our experiments [seq\_data, phenoTable] fastq => bam => gene-cell count table; phenotype table.
      2. data imported from other resources [GEO]
   2. Define inputs, intermediates and outputs
      1. Intermediates [step\_wise\_data, data imported from other resources]
      2. Outputs [figures, tables]
      3. Start from different inputs combination [version control]
   3. Storage and transfer
      1. Google drive
      2. Local drive
2. **Pipeline code organization**
   1. One task, one module [title, inputs, packages, outputs]: module size?
   2. Analysis platform?
      1. Python => pattern recognition
      2. R => Statistics
   3. Data manipulation?
      1. Analysis platform conversion (Python ⬄ R: Data format conversion <-> different data storage formats)
      2. Data frame annotation
      3. Code sharing and version control
   4. **Problem solving**

Communicate with the original authors to solve technical issues

Identification of a better solution with benchmarks

* 1. Used by a wet-lab user

1. **Actual data processing pipelines**

**Goals: identify and visualize the patterns that are biologically meaningful**

**Identify the significant disturbance of the patterns in patients**

1. Understand and optimize an established basic scRNAseq analysis pipeline
2. Preprocess

=> (francesco\_aggr\_pasca.ipynb): Aggregation to avoid bias, an ‘integrated’ dataset for downstream analysis; Data QC & cleaning up

=> Data objects conversion between R and Python platforms

Inputs:  
Outputs:

1. (francesco\_1\_pasca.ipynb):

Data Integration (Liger) => Integrating Multiple scRNA-seq Datasets

Inputs:  
Outputs:

* <https://github.com/welch-lab/liger> (Integrating Multiple Single-Cell RNA-seq Datasets)
* Benchmarking ref: https://www.nature.com/articles/s41592-021-01336-8

Data annotation (SingleR) => knowledge mapping

Inputs:  
Outputs:

<https://www.bioconductor.org/packages/release/bioc/vignettes/SingleR/inst/doc/SingleR.html>

1. (seurat\_cc\_deg\_calc.Rmd): **Feature analysis**:

Inputs:

Outputs:

(<https://satijalab.org/seurat/articles/get_started.html>) => statistical approach is preferred (FDR p value, fold changes, sample size, or proper measurement)

* 1. Feature selection
  2. Scaling the data
  3. Linear dimensional reduction (PCA)
  4. Determine the ‘dimensionality’ of the dataset
  5. Clustering
  6. non-linear dimensional reduction (UMAP)
  7. Find and visual any differentially expressed features (cluster or other annotation specific biomarkers)
  8. Perform DE analysis using alternative tests (MAST and DESeq2) => alterations in disease condition

**Deliverables:**

1. a modularized and standardized analysis pipeline that can be performed by a wet-lab user.
2. Altered cell types and Altered expressed gene table with statistics in 22q11DS patients.
3. Advanced analysis pipelines

The idea: Integration of spatial transcriptomics in single-cell study

(<https://www.sciencedirect.com/science/article/pii/S1672022922000845>)

* + spatial landmarks for every cell–region combination, can predict multiple likely positions for each cell (Seurat)
  + a spatial signature genes (zipcodes) mapping (Geo-seq)
  + neighboring cells display more similar transcriptional identities than cells farther apart and the physical distance between cells increases with their biological distance in gene expression space. (novoSpaRc)
  + ligand–receptor interaction information encodes the cellular spatial organization (CSOmap, CellPhoneDB3.0)

1. **Integrated intra- and intercellular signaling knowledge analysis (spatial analysis)**
   1. intercellular interactions and communication

Software & database: CellPhoneDB

(<https://www.cellphonedb.org/explore-sc-rna-seq>)

Trend review: (https://www.nature.com/articles/s41576-020-00292-x)

**Deliverables:**

1. an intercellular network analysis pipeline.
2. Altered intercellular-signalings in 22q11DS.
3. **Integrating single-cell and spatial transcriptomics to elucidate tissue structure and dynamics (spatial information integration)**
   1. Tangram

Software: <https://github.com/broadinstitute/Tangram>

(<https://pubmed.ncbi.nlm.nih.gov/31178122/>)

Trend review:

(<https://www.nature.com/articles/s41592-022-01480-9>)

Ref: https://github.com/QuKunLab/SpatialBenchmarking

(<https://www.cell.com/cell/pdf/S0092-8674(19)30559-8.pdf)>

(https://www.sciencedirect.com/science/article/pii/S1672022922000845#s0060)

**Deliverables:**

1. Spatial info integration analysis pipeline.
2. Potential altered brain region/layers in 22q11DS brain

Additional thoughts:

1. **Integrated intra- and intercellular signaling knowledge analysis (**temporal analysis)

RNA velocity analysis by cellrank

Software: (<https://pubmed.ncbi.nlm.nih.gov/31178122/>)

Ref data:

Trend Review:

Trajectory Benchmarking https://www.nature.com/articles/s41587-019-0071-9

(https://www.sciencedirect.com/science/article/pii/S0958166919301430)

**Deliverables:**

1. Cell lineage analysis pipeline.
2. Identify potential altered cell lineages in 22q11DS