

Data exploratory analysis

Antonia Chroni for SJCRH DNB_BINF_Core

Contents

1	Information about this notebook	3
2	Set up	3
3	Directories and paths to file Inputs/Outputs	3
4	Read raw data file	4
5	Read processed data file	4
6	Update processed data file per our conversation with Jackie Norrie	4
7	10x matched scRNA-seq and scATAC-seq cohort	5
7.1	Type of sequencing assay and unit	5
7.2	Number of samples per experiment	6
7.3	Number of samples per seq_technology_assay	7
7.4	Summary of samples	8
8	10x Genomics Multiome cohort	9
8.1	Type of sequencing assay and seq_unit	9
8.2	Number of samples per experiment	9
8.3	Number of samples per seq_technology_assay	10
8.4	Summary of samples	11
9	Notes	12
10	References	12
11	Session Info	13
## The following object is masked _by_ .GlobalEnv:		
##		
## root_dir		

PI: NA

Project: test-dataset

Task: NA

Project Lead(s): NA

Department: Developmental Neurobiology

DNB Bioinformatics Core Analysis Team:

Lead Analyst(s): Antonia Chroni, PhD

Group Lead: Cody A. Ramirez, PhD

Contact E-mail: antonia.chroni@stjude.org¹

DNB Bioinformatics Core Pipeline: Standard sc-/sn-ATAC-Seq Analysis in 10X Genomics data

Date started: Dec-16-2024

Date completed: ONGOING

Report generated: 15:16:20 CST 01/09/2025

Reviewed by: _____ Date: _____

¹<mailto:antonia.chroni@stjude.org>

1 Information about this notebook

This is an exploratory analysis of the data available for the testing phase for the **sc-atac-seq** pipeline(s). We are investigating number of samples overall per condition and variables as defined in the **params**. We are looking for cohorts that fit the following criteria as described in the `./analyses/README.md`². - Control vs condition (min. 3+3 samples) - Number of cells/sample - size of datasets might determined packages/pipelines to be used - Single cell/Single-nucleus ATAC - Integration of scRNA-seq and scATAC-seq data from the same biological system (multiple modalities) - Pipeline for same samples but not same cells with scRNA-seq and scATAC-seq - Annotate scATAC-seq cells via label transfer by using scRNA data: cell type annotation - Available bulk ATAC-seq data for the same samples (matched) - this could be used for cell type annotation

2 Set up

```
suppressPackageStartupMessages({  
  library(tidyverse)  
  library(knitr)  
  library(readxl)  
})
```

3 Directories and paths to file Inputs/Outputs

```
attach(params)  
  
## The following object is masked _by_ .GlobalEnv:  
##  
##      root_dir  
  
analysis_dir <- file.path(root_dir, "analyses", "data-exploratory-analysis")  
  
# input files  
data_file <- file.path(metadata_dir, "TestData_10x_2024_12_16.xlsx")  
jackie_data_file <- file.path(metadata_dir, "cohorts_10x_rna_atac_testing_phase_JN.xlsx")  
  
# File path to `input` directory  
input_dir <- file.path(analysis_dir, "input")  
if (!dir.exists(input_dir)) {  
  dir.create(input_dir)}  
  
# File path to `plots` directory  
plots_dir <- file.path(analysis_dir, "plots")  
if (!dir.exists(plots_dir)) {  
  dir.create(plots_dir)}  
  
# File path to `results` directory  
results_dir <- file.path(analysis_dir, "results")  
if (!dir.exists(results_dir)) {  
  dir.create(results_dir)}
```

²<https://github.com/stjude-dnb-binfcore/sc-atac-seq/tree/main/analyses>

4 Read raw data file

```
# Read metadata
raw_data_df <- read_excel(data_file) %>%

# save data under a more descriptive file name
# all edits will be made on this one and not on the RAW data
# RAW data were edited manually by Antonia Chroni to add typo of experiment and type of k
write_tsv(file.path(input_dir, "cohorts_10x_rna_atac_testing_phase_not_processed.tsv"))
```

5 Read processed data file

```
df_processed <- raw_data_df %>%

# Add metadata
add_column(seq_unit = "nucleus",
           condition = "unknown",
           species = "mouse") %>%
tidyr::separate(Kit, c("seq_technology", "assay_drop"), sep = ' ', remove = FALSE) %>%
tidyr::separate(assay_drop, c("assay", "drop"), sep = ' ', remove = FALSE) %>%
select(!c(assay_drop, drop)) %>%
mutate(seq_unit = case_when(grepl("More Retina", experiment) ~ "cell",
                           grepl("Stressed Retina", experiment) ~ "cell",
                           TRUE ~ seq_unit),
       Sample = case_when(grepl("6wk Cerebellum", Sample) ~ "6 week cerebellum",
                           grepl("P0 Cerebellum,", Sample) ~ "P0 cerebellum",
                           grepl("E14.5", Sample) ~ "E14.5 Retina",
                           TRUE ~ Sample),

# Add condition col
condition = case_when(grepl("P0 cerebellum|6 week cerebellum|E14.5|P0 Retina|E14.5",
                           grepl("Wt1|Wt2|SEKO_21|SEKO_25", Sample) ~ "knock-out",
                           grepl("ATOH HI|ATOH Low", Sample) ~ "treatment",
                           TRUE ~ condition)) %>%

write_tsv(file.path(results_dir, "cohorts_10x_rna_atac_testing_phase_2024-12-17.tsv"))
```

6 Update processed data file per our conversation with Jackie Norrie

This is the cohorts_10x_rna_atac_testing_phase_2024-12-17.tsv file updated per our conversation with Jackie Norrie to fix empty cells and/or inconsistencies.

```
df_processed_select <- df_processed %>%
  select(!c(SRM_id, SRM_Sample_id, seq_unit, condition))

# Read and process data
df_processed <- read_excel(jackie_data_file) %>%
  select(DYE, SRM_id, SRM_Sample_id, seq_unit, condition) %>%
  right_join(df_processed_select, by = join_by(DYE)) %>%
```



```

# What type of assay was used?
assay_samples <- unique(df$assay)

# What type of seq_technology was used?
seq_technology_assay_samples <- unique(df$seq_technology_assay)

# What type of seq_technology was used?
PI_samples <- unique(df$PI)

```

Single cell sequencing was done by using nucleus, cell and there are 10Xv2_ATAC, 10Xv3_RNA sequencing technologies and assays in the database. Samples are generated by the Dyer lab. Cohort is formed and processed accordingly.

7.2 Number of samples per experiment

Table 1: Summary of samples per experiment

experiment	n
Cerebellum	9
Stressed Retina	3
Victoria Knockout	8

7.3 Number of samples per seq_technology_assay

Here, we investigate the number of libraries per assay, i.e., `seq_technology_assay` and per `matched_samples_ID`.

Table 2: Number of samples per seq_technology_assay

experiment	tissue	condition	matched_samples_ID	seq_unit	seq_technology_assay	n
Cerebellum	Cerebellum	age	6 week cerebellum_1	cell	10Xv3_RNA	1
Cerebellum	Cerebellum	age	6 week cerebellum_1	nucleus	10Xv2_ATAC	2
Cerebellum	Cerebellum	age	P0 cerebellum_10	cell	10Xv3_RNA	1
Cerebellum	Cerebellum	age	P0 cerebellum_10	nucleus	10Xv2_ATAC	1
Cerebellum	Cerebellum	sorted	ATOH HI_2	cell	10Xv3_RNA	1
Cerebellum	Cerebellum	sorted	ATOH HI_2	nucleus	10Xv2_ATAC	1
Cerebellum	Cerebellum	sorted	ATOH Low_3	cell	10Xv3_RNA	1
Cerebellum	Cerebellum	sorted	ATOH Low_3	nucleus	10Xv2_ATAC	1
Stressed Retina	Retina	LPS Injection	LPS_6	cell	10Xv3_RNA	1
Stressed Retina	Retina	LPS Injection	LPS_ATAC_7	nucleus	10Xv2_ATAC	1
Stressed Retina	Retina	PBS Injection	PBS_11	cell	10Xv3_RNA	1
Victoria Knockout	Retina	knock-out	SEKO_21_13	cell	10Xv3_RNA	1
Victoria Knockout	Retina	knock-out	SEKO_21_13	nucleus	10Xv2_ATAC	1
Victoria Knockout	Retina	knock-out	SEKO_25_14	cell	10Xv3_RNA	1
Victoria Knockout	Retina	knock-out	SEKO_25_14	nucleus	10Xv2_ATAC	1
Victoria Knockout	Retina	wt	Wt1_15	cell	10Xv3_RNA	1
Victoria Knockout	Retina	wt	Wt1_15	nucleus	10Xv2_ATAC	1
Victoria Knockout	Retina	wt	Wt2_16	cell	10Xv3_RNA	1
Victoria Knockout	Retina	wt	Wt2_16	nucleus	10Xv2_ATAC	1

7.4 Summary of samples

Table 3: Summary of samples

experiment	tissue	condition	matched_samples_ID	seq_unit	seq_technology_assay
Cerebellum	Cerebellum	age	6 week cerebellum_1	nucleus	10Xv2_ATAC
Cerebellum	Cerebellum	age	6 week cerebellum_1	nucleus	10Xv2_ATAC
Cerebellum	Cerebellum	age	6 week cerebellum_1	cell	10Xv3_RNA
Cerebellum	Cerebellum	age	P0 cerebellum_10	nucleus	10Xv2_ATAC
Cerebellum	Cerebellum	age	P0 cerebellum_10	cell	10Xv3_RNA
Cerebellum	Cerebellum	sorted	ATOH HI_2	nucleus	10Xv2_ATAC
Cerebellum	Cerebellum	sorted	ATOH HI_2	cell	10Xv3_RNA
Cerebellum	Cerebellum	sorted	ATOH Low_3	nucleus	10Xv2_ATAC
Cerebellum	Cerebellum	sorted	ATOH Low_3	cell	10Xv3_RNA
Stressed Retina	Retina	LPS Injection	LPS_6	cell	10Xv3_RNA
Stressed Retina	Retina	LPS Injection	LPS_ATAC_7	nucleus	10Xv2_ATAC
Stressed Retina	Retina	PBS Injection	PBS_11	cell	10Xv3_RNA
Victoria Knockout	Retina	knock-out	SEKO_21_13	nucleus	10Xv2_ATAC
Victoria Knockout	Retina	knock-out	SEKO_21_13	cell	10Xv3_RNA
Victoria Knockout	Retina	knock-out	SEKO_25_14	nucleus	10Xv2_ATAC
Victoria Knockout	Retina	knock-out	SEKO_25_14	cell	10Xv3_RNA
Victoria Knockout	Retina	wt	Wt1_15	nucleus	10Xv2_ATAC
Victoria Knockout	Retina	wt	Wt1_15	cell	10Xv3_RNA
Victoria Knockout	Retina	wt	Wt2_16	nucleus	10Xv2_ATAC
Victoria Knockout	Retina	wt	Wt2_16	cell	10Xv3_RNA

8 10x Genomics Multiome cohort

```
df <- df_processed %>%  
  filter(multiome_10x == "yes")
```

8.1 Type of sequencing assay and seq_unit

We should investigate if there are samples from two different sequencing technologies and unit.

```
# Was nucleus or whole cell used for the sequencing?  
seq_unit_samples <- unique(df$seq_unit)  
  
# What type of assay was used?  
assay_samples <- unique(df$assay)  
  
# What type of seq_technology was used?  
seq_technology_assay_samples <- unique(df$seq_technology_assay)  
  
# What type of seq_technology was used?  
PI_samples <- unique(df$PI)
```

Single cell sequencing was done by using nucleus and there are 10X_ATAC, 10X_RNA sequencing technologies and assays in the database. Samples are generated by the Dyer lab. Cohort is formed and processed accordingly.

8.2 Number of samples per experiment

Table 4: Summary of samples per experiment

experiment	n
Multiome Retina	6

8.3 Number of samples per seq_technology_assay

Here, we investigate the number of libraries per assay, i.e., `seq_technology_assay` and per `matched_samples_ID`.

Table 5: Number of samples per seq_technology_assay

experiment	tissue	condition	matched_samples_ID	seq_unit	seq_technology_assay	n
Multiome Retina	Retina	age	Adult Retina_4	nucleus	10X_ATAC	1
Multiome Retina	Retina	age	Adult Retina_4	nucleus	10X_RNA	1
Multiome Retina	Retina	age	E14.5 Retina_5	nucleus	10X_ATAC	1
Multiome Retina	Retina	age	E14.5 Retina_5	nucleus	10X_RNA	1
Multiome Retina	Retina	age	P0 Retina_9	nucleus	10X_ATAC	1
Multiome Retina	Retina	age	P0 Retina_9	nucleus	10X_RNA	1

8.4 Summary of samples

Table 6: Summary of samples

experiment	tissue	condition	matched_samples_ID	seq_unit	seq_technology_assay
Multiome Retina	Retina	age	Adult Retina_4	nucleus	10X_ATAC
Multiome Retina	Retina	age	Adult Retina_4	nucleus	10X_RNA
Multiome Retina	Retina	age	E14.5 Retina_5	nucleus	10X_ATAC
Multiome Retina	Retina	age	E14.5 Retina_5	nucleus	10X_RNA
Multiome Retina	Retina	age	P0 Retina_9	nucleus	10X_ATAC
Multiome Retina	Retina	age	P0 Retina_9	nucleus	10X_RNA

9 Notes

I have identified the following datasets that **almost** fit the criteria for the testing phase:

- **10x matched scRNA-seq and scATAC-seq:** **Victoria Knockout** experiment with 1 replicate/knock-out group (4 samples for 10x RNA + 4 samples 10x ATAC). Replicates could be potentially grouped together and have 2 replicates for WT and 2 replicates for knock-out.
- **10x Genomics Multiome:** **Multiome Retina** experiment with 1 replicate/age group (3 samples for 10x RNA (Multiome) + 3 samples for 10x ATAC (multiome)).

10 References

- **More Retina and Stressed retina** experiments published by Norrie et al., 2025³.
- **Victoria Knockout** experiment published by Honnell et al., 2022⁴.

³<https://doi.org/10.1016/j.devcel.2024.12.014>

⁴<https://www.nature.com/articles/s41467-021-27924-y#Sec13>

11 Session Info

```
## R version 4.4.0 (2024-04-24)
## Platform: x86_64-pc-linux-gnu
## Running under: Red Hat Enterprise Linux 8.8 (Ootpa)
##
## Matrix products: default
## BLAS: /usr/lib64/libblas.so.3.8.0
## LAPACK: /usr/lib64/liblapack.so.3.8.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8      LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8  LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8     LC_NAME=C
##  [9] LC_ADDRESS=C             LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## time zone: America/Chicago
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods    base
##
## other attached packages:
##  [1] readxl_1.4.3    knitr_1.48      lubridate_1.9.3 forcats_1.0.0
##  [5] stringr_1.5.1   dplyr_1.1.4     purrr_1.0.2     readr_2.1.5
##  [9] tidyr_1.3.1     tibble_3.2.1    ggplot2_3.5.1   tidyverse_2.0.0
## [13] yaml_2.3.10
##
## loaded via a namespace (and not attached):
##  [1] bit_4.0.5      gtable_0.3.5    jsonlite_1.8.8   crayon_1.5.3
##  [5] compiler_4.4.0 tidyselect_1.2.1 parallel_4.4.0    jquerylib_0.1.4
##  [9] scales_1.3.0   fastmap_1.2.0   mime_0.12        R6_2.5.1
## [13] generics_0.1.3 munsell_0.5.1   bslib_0.8.0      pillar_1.9.0
## [17] tzdb_0.4.0     rlang_1.1.4     utf8_1.2.4       stringi_1.8.4
## [21] cachem_1.1.0   xfun_0.47       sass_0.4.9       bit64_4.0.5
## [25] timechange_0.3.0 cli_3.6.3       withr_3.0.1      magrittr_2.0.3
## [29] digest_0.6.37  grid_4.4.0      vroom_1.6.5      hms_1.1.3
## [33] lifecycle_1.0.4 vctrs_0.6.5     evaluate_0.24.0   glue_1.7.0
## [37] cellranger_1.1.0 fansi_1.0.6     colorspace_2.1-1  rmarkdown_2.28
## [41] tools_4.4.0    pkgconfig_2.0.3 htmltools_0.5.8.1
```