Data exploratory analysis

Antonia Chroni for SJCRH DNB_BINF_Core

Contents


```
## 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.org[1]
DNB Bioinformatics Core Pipeline: Standard sc-/sn-ATAC-Seq Analysis in 10X Genomics data

Date started: Dec-16-2024
Date completed: ONGOING
Report generated: 12:35:53 CST 01/03/2025

Reviewed by: _____ Date: _____

[1] mailto:antonia.chroni@stjude.org

1 Information about this notebook

This is an exploratory analysis of the data in the project. 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.

- 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

3 Directories and paths to file Inputs/Outputs

4 Read raw data file

5 Read processed data file

5.1 Color palette for plotting

```
# Read color palette
#palette_df <- readr::read_tsv(palette_file, guess_max = 100000, show_col_types = FALSE) %>%
#  mutate(color_names = case_when(grepl("binary_1", color_names) ~ "Het",
#                                 grepl("binary_2", color_names) ~ "WT"))

# Define and order palette
#palette <- palette_df$hex_codes
#names(palette) <- palette_df$color_names
```

6 10x matched scRNA-seq and scATAC-seq cohort

6.1 Type of sequencing assay and unit

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

Single cell sequencing was done by using nucleus, cell and there are RNA, ATAC and 10Xv3, 10Xv2 sequencing technologies in the database. Cohort is formed and processed accordingly.

6.2 Number of samples per experiment

Table 1: Summary of samples per experiment

experiment	n
Cerebellum	9
More Retina	7

experiment	n
Stressed Retina	3
Victoria Knockout	8

6.3 Number of samples per experiment, seq_unit, assay, and condition

Table 2: Summary of samples per experiment, seq_unit, assay, and condition

experiment	seq_unit	assay	condition	n
Cerebellum	nucleus	ATAC	age	3
Cerebellum	nucleus	ATAC	treatment	2
Cerebellum	nucleus	RNA	age	2
Cerebellum	nucleus	RNA	treatment	2
More Retina	cell	ATAC	age	2
More Retina	cell	ATAC	unknown	3
More Retina	cell	RNA	age	1
More Retina	cell	RNA	unknown	1
Stressed Retina	cell	ATAC	unknown	1
Stressed Retina	cell	RNA	unknown	2
Victoria Knockout	nucleus	ATAC	knock-out	4
Victoria Knockout	nucleus	RNA	knock-out	4

6.4 Number of samples per experiment, seq_unit, assay, condition, and sample

Table 3: Summary of samples per experiment, seq_unit, assay, seq_technology, condition, and Sample

experiment	seq_unit	assay	seq_technology	condition	Sample	n
Cerebellum	nucleus	ATAC	10Xv2	age	6 week cerebellum	2
Cerebellum	nucleus	ATAC	10Xv2	age	P0 cerebellum	1
Cerebellum	nucleus	ATAC	10Xv2	treatment	ATOH HI	1
Cerebellum	nucleus	ATAC	10Xv2	treatment	ATOH Low	1
Cerebellum	nucleus	RNA	10Xv3	age	6 week cerebellum	1
Cerebellum	nucleus	RNA	10Xv3	age	P0 cerebellum	1
Cerebellum	nucleus	RNA	10Xv3	treatment	ATOH HI	1
Cerebellum	nucleus	RNA	10Xv3	treatment	ATOH Low	1
More Retina	cell	ATAC	10Xv2	age	E14.5 Retina	2
More Retina	cell	ATAC	10Xv2	unknown	NRL Dep	1
More Retina	cell	ATAC	10Xv2	unknown	Retina	2
More Retina	cell	RNA	10Xv3	age	E14.5	1
More Retina	cell	RNA	10Xv3	unknown	NRL Dep	1
Stressed Retina	cell	ATAC	10Xv2	unknown	LPS_ATAC	1
Stressed Retina	cell	RNA	10Xv3	unknown	LPS	1
Stressed Retina	cell	RNA	10Xv3	unknown	PBS	1
Victoria Knockout	nucleus	ATAC	10Xv2	knock-out	SEKO_21	1
Victoria Knockout	nucleus	ATAC	10Xv2	knock-out	SEKO_25	1
Victoria Knockout	nucleus	ATAC	10Xv2	knock-out	Wt1	1
Victoria Knockout	nucleus	ATAC	10Xv2	knock-out	Wt2	1
Victoria Knockout	nucleus	RNA	10Xv3	knock-out	SEKO_21	1
Victoria Knockout	nucleus	RNA	10Xv3	knock-out	SEKO_25	1
Victoria Knockout	nucleus	RNA	10Xv3	knock-out	Wt1	1
Victoria Knockout	nucleus	RNA	10Xv3	knock-out	Wt2	1

7 10x Genomics Multiome cohort

7.1 Type of sequencing assay and seq_unit

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

Single cell sequencing was done by using nucleus and there are RNA, ATAC and 10Xv3, 10Xv2 sequencing technologies in the database. Cohort is formed and processed accordingly.

7.2 Number of samples per experiment

Table 4: Summary of samples per experiment

experiment	n
Multiome Retina	6

7.3 Number of samples per experiment, seq_unit, assay, and condition

Table 5: Summary of samples per experiment, seq_unit, assay, and condition

experiment	seq_unit	assay	condition	n
Multiome Retina	nucleus	ATAC	age	3
Multiome Retina	nucleus	RNA	age	3

7.4 Number of samples per experiment, seq_unit, assay, seq_technology, condition, and sample

Table 6: Summary of samples per experiment, seq_unit, assay, seq_technology, condition, and Sample

experiment	seq_unit	assay	seq_technology	condition	Sample	n
Multiome Retina	nucleus	ATAC	10Xv2	age	Adult Retina	1
Multiome Retina	nucleus	ATAC	10Xv2	age	E14.5 Retina	1
Multiome Retina	nucleus	ATAC	10Xv2	age	P0 Retina	1
Multiome Retina	nucleus	RNA	10Xv3	age	Adult Retina	1
Multiome Retina	nucleus	RNA	10Xv3	age	E14.5 Retina	1
Multiome Retina	nucleus	RNA	10Xv3	age	P0 Retina	1

8 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).

9 Questions

1. Are samples in the `seq_unit` column correctly assigned?
2. Are samples in the `condition` column correctly assigned?
3. What is the column to use for `sample_id`? Is the `DYE` column? That will help me to summarize and confirm samples with paired assays per experiment.

10 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
```