Study of the regulatory programs of spermatogenesis through the integration of single-cell RNA and ATAC

Projects in Bioinformatics - Fall 2025

Johan Olesen
202104408
Msc. Student Bioinformatics

Samuele Soraggi
Supervisor
Special consultant, Bioinformatics Research Center, Aarhus University

09.01.2026

Contents

1 Introduction

Spermatogenesis is a complex process that permits the differentiation of stem cells into mature spermatozoa, and is of high relevance in studying infertility conditions and cross-species differences in the biological processes.

1.1 Goals for the project

Initial:

- learn basics of git
- learn sc workflow with scanpy, muon and scvi-tools
- work with real messy data
- Answer:
  - Cell states & trajectories: Can we recover a clean spermatogenic trajectory (spermatogonia → spermatocytes → spermatids) and supporting somatic lineages?
  - Peak→gene linkage: Which distal elements likely regulate stage-specific genes?
  - TF programs: Which TFs show coordinated motif accessibility + target expression? (e.g., STRA8, A-MYB, TAF7L)

Actually done:

- learn basics of git
- learn sc workflow with scanpy, muon and scvi-tools
- work with real messy data
- Answer: Cell states & trajectories: Can we recover a clean spermatogenic trajectory (spermatogonia → spermatocytes → spermatids) and supporting somatic lineages?
- Celltype annotation of both scRNA-seq and scATAC-seq.
- Cell topic for scATAC-seq

1.2 Workflow Overview

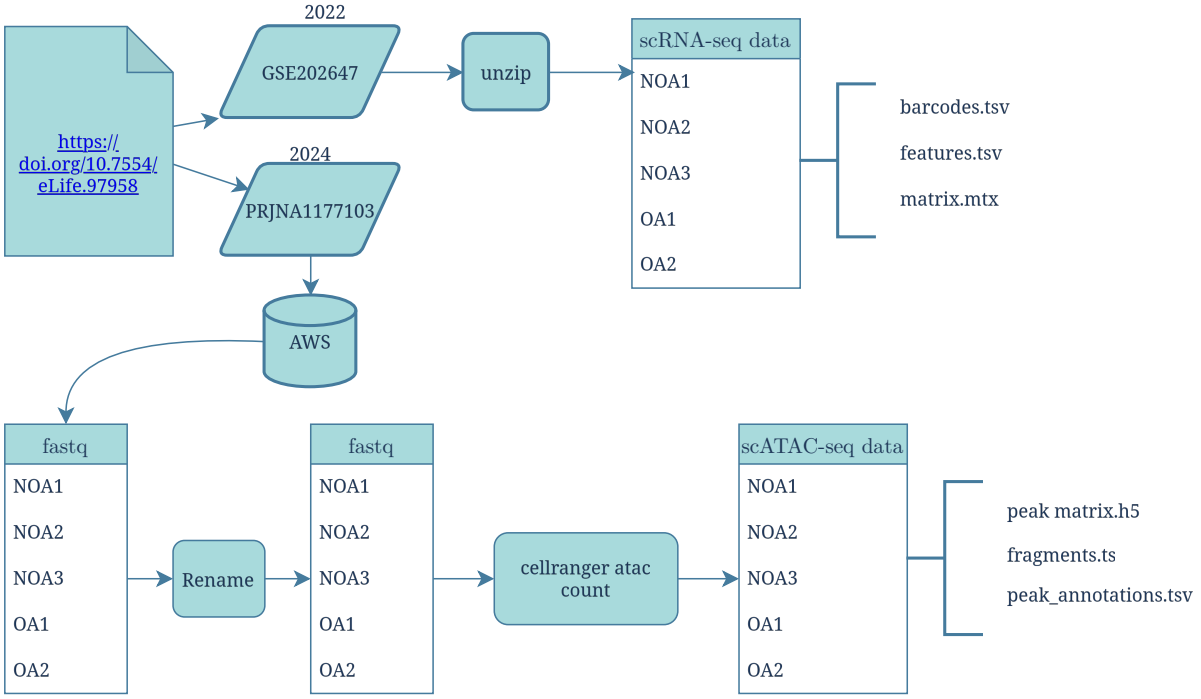

Figure 1: Stage 1 schematic of data aquisition and preparation.

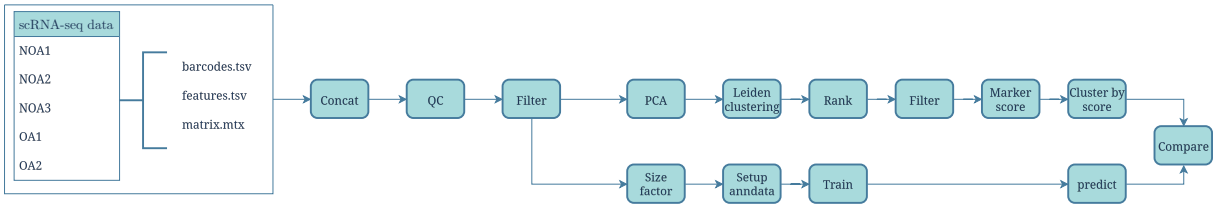

Figure 2: Stage 2 schematic of scRNA-seq celltype annotation with a *semi-manual* way and CellAssign model.

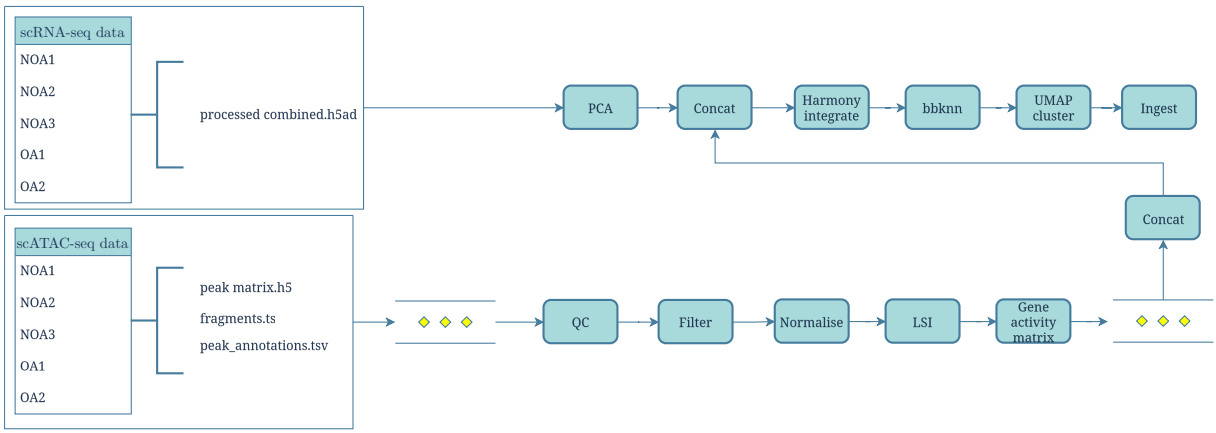

Figure 3: Stage 3 schematic of attempt at label transfer by integrating scRNA-seq and scATAC-seq.

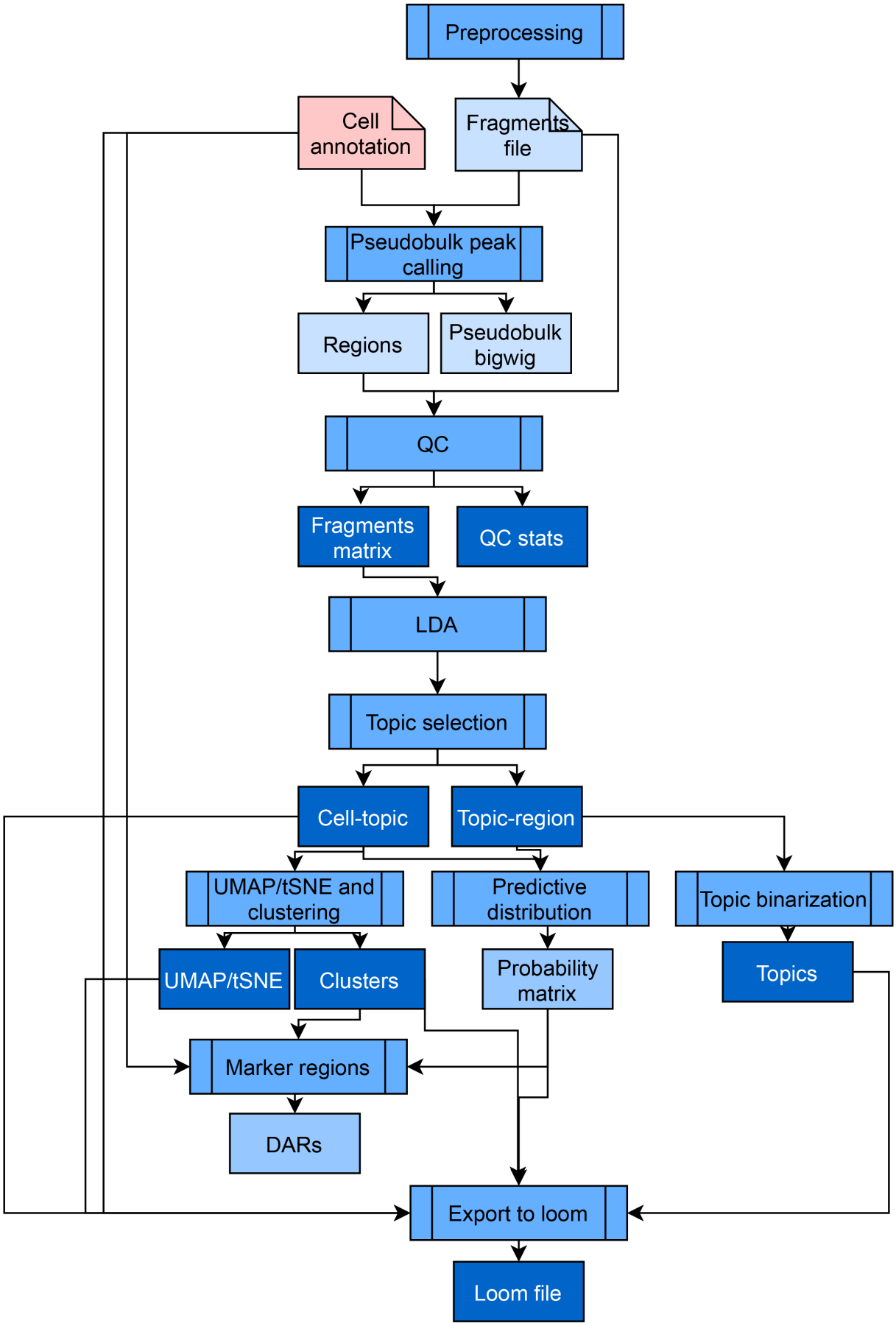

Figure 4: Stage 3 schematic of pycistopic workflow from PyCistopic documentation. [1]

Flowcharts drawn with `draw.io`.

1.2.1 Environment setup with Conda

First step was to get a working environment setup for the analyses. For this Conda was used to create a environment with the required packages, relying on `pip` for the most up-to-date packages.

For the tutorial run and scRNA-seq labelling the environmnet `torch_env.yml` [2] was used. This environment includes the scverse's anndata [3], mudata [4], scanpy [5], muon [6] and scvi-tools [7] packages, as well as full PyTorch [8] CUDA capabilities for scvi-tools.

For the second scATAC-seq workflow another environment was used because of versioning requirements; `cistopic_env.yml`, consisting of the SCENIC+ [9] suite.

1.2.2 Tutorial run

To start of before real testis data had been found. A quick run through of the tutorial run of multiome 10X PBMC [10] by Bredikin to get a quick overview of how to work with single cell data and anndata objects, and to check Conda environment worked.

Was succesful in creating the same analysis as the tutorial.

2 Stage 1: Data aquisition and preparation

For the first stage of the project, we will be focusing on acquiring and preparing data for analysis. This includes downloading and organizing the necessary datasets, as well as preprocessing the data to ensure it is ready for downstream analysis. The workflow is illustrated on **Figure 1**

For the real dataset set out in the goals, Wang et al [11] have made their scRNA-seq and scATAC-seq data available.

The scRNA-seq data was available under NCBI Gene Expression Omnibus ID **GSE202647** from 2022, and was already ready for analysis.

The scATAC-seq data was available under NCBI BioProject ID **PRJNA1177103** from 2024. This was only the raw read data, so the Cellranger ATAC [12] pipeline was run for each of the five samples. For Cellranger to be able to run the data had to be structured and named in a specific way. Using the guidelines available from the offical 10X Cellranger documentation and matching the read length of each of the four files per sample was renamed accordingly.

The `count` function was run with Cellranger ATAC version 2.2.0 on each sample with reference data *refdata-cellranger-arc-GRCh38-2024-A* using SLURM. This resulted in the `peak_matrix.h5`, `fragments.tsv` and `peak_annotations.tsv` for each sample ready for analysis.

Ideally, we wanted the data from scRNA-seq and scATAC-seq to be cell matched from their barcodes, but finding public datasets for testis with that criteria was difficult. Instead we will be matching cells by celltype instead. The data in Wang et al's study comes from the same five donors, but sequenced at seperate time points.

3 Stage 2: Celltype annotation of scRNA-seq data

4 Stage 3a: Celltype annotation of scATAC-seq data with label transfer

5 Stage 3b: Celltype annotation of scATAC-seq data with pycistopic

6 Conclusion

References

[1] González-Blas CB, Hulselmans G. Features — pycisTopic Documentation. [accessed 2026 Jan 9]. https://pycistopic.readthedocs.io/en/latest/features.html

[2] Soraggi S. SamueleSoraggi/PIB-johan-olesen. 2025 Dec 18 [accessed 2026 Jan 9]. https://github.com/SamueleSoraggi/PIB-johan-olesen

[3] Virshup I, Rybakov S, Theis FJ, Angerer P, Wolf FA. Anndata: Access and Store Annotated Data Matrices. Journal of Open Source Software. 2024 [accessed 2026 Jan 9];9(101):4371. https://joss.theoj.org/papers/10.21105/joss.04371. doi:10.21105/joss.04371

[4] Virshup I, Bredikhin D, Heumos L, Palla G, Sturm G, Gayoso A, Kats I, Koutrouli M, Berger B, Pe'er D, et al. The Scverse Project Provides a Computational Ecosystem for Single-Cell Omics Data Analysis. Nature Biotechnology. 2023 [accessed 2026 Jan 9];41(5):604–606. https://www.nature.com/articles/s41587-023-01733-8. doi:10.1038/s41587-023-01733-8

[5] Wolf FA, Angerer P, Theis FJ. SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis. Genome Biology. 2018 [accessed 2026 Jan 9];19(1):15. https://doi.org/10.1186/s13059-017-1382-0. doi:10.1186/s13059-017-1382-0

[6] Bredikhin D, Kats I, Stegle O. MUON: Multimodal Omics Analysis Framework. Genome Biology. 2022 [accessed 2026 Jan 9];23(1):42. https://doi.org/10.1186/s13059-021-02577-8. doi:10.1186/s13059-021-02577-8

[7] Gayoso A, Lopez R, Xing G, Boyeau P, Valiollah Pour Amiri V, Hong J, Wu K, Jayasuriya M, Mehlman E, Langevin M, et al. A Python Library for Probabilistic Analysis of Single-Cell Omics Data. Nature Biotechnology. 2022 [accessed 2026 Jan 9];40(2):163–166. https://www.nature.com/articles/s41587-021-01206-w. doi:10.1038/s41587-021-01206-w

[8] Ansel J, https://orcid.org/0009-0007-5207-2179, View Profile, Yang E, https://orcid.org/0009-0008-0621-7872, View Profile, He H, https://orcid.org/0009-0004-1133-816X, View Profile, Gimelshein N, et al. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2. 2024 [accessed 2026 Jan 9]:929–947. (ACM Conferences). https://dlnext.acm.org/doi/10.1145/3620665.3640366. doi:10.1145/3620665.3640366

[9] Bravo González-Blas C, De Winter S, Hulselmans G, Hecker N, Matetovici I, Christiaens V, Poovathingal S, Wouters J, Aibar S, Aerts S. SCENIC+: Single-Cell Multiomic Inference of Enhancers and Gene Regulatory Networks. Nature Methods. 2023 [accessed 2026 Jan 9];20(9):1355–1367. https://www.nature.com/articles/s41592-023-01938-4. doi:10.1038/s41592-023-01938-4

[10] Bredikhin D. Processing Gene Expression of 10k PBMCs — Muon-Tutorials Documentation. [accessed 2026 Jan 9]. https://muon-tutorials.readthedocs.io/en/latest/single-cell-rna-atac/pbmc10k/1-Gene-Expression-Processing.html

[11] Wang S, Wang H, Jin B, Yan H, Zheng Q, Zhao D. scRNA-seq and scATAC-seq Reveal That Sertoli Cell Mediates Spermatogenesis Disorders through Stage-Specific Communications in Non-Obstructive Azoospermia Park J, Choi M, editors. eLife. 2025 [accessed 2025 Sept 12];13:RP97958. https://doi.org/10.7554/eLife.97958. doi:10.7554/eLife.97958

[12] Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, Olsen BN, Mumbach MR, Pierce SE, Corces MR, et al. Massively Parallel Single-Cell Chromatin Landscapes of Human Immune Cell Development and Intratumoral T Cell Exhaustion. Nature Biotechnology. 2019 [accessed 2026 Jan 9];37(8):925–936. https://www.nature.com/articles/s41587-019-0206-z. doi:10.1038/s41587-019-0206-z

[13] Specifying Input FASTQ Files for Cell Ranger ATAC | Official 10x Genomics Support. [accessed 2026 Jan 9]. https://www.10xgenomics.com/support/software/cell-ranger-atac/latest/analysis/inputs/specifying-input-fastq-files

Appendix A Cellranger

Naming scheme for Cellranger ATAC count [13]: `[Sample Name]S1_L00[Lane Number][Read Type]_001.fastq.gz`, where `Read type`:

- `I1`: Dual index i7 read (optional)
- `R1`: Read 1
- `I2`: Dual index i5 read
- `R3`: Read 2

`Lane Number` does not matter. `Sample Name` can be anything.

Example of NOA1 sample:

- SRR31097965_S1_L001_I1_001.fastq
- SRR31097965_S1_L001_I2_001.fastq
- SRR31097965_S1_L001_R1_001.fastq
- SRR31097965_S1_L001_R2_001.fastq

Appendix B scRNA-seq

Appendix C scATAC-seq

Appendix D pycistopic workflow