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

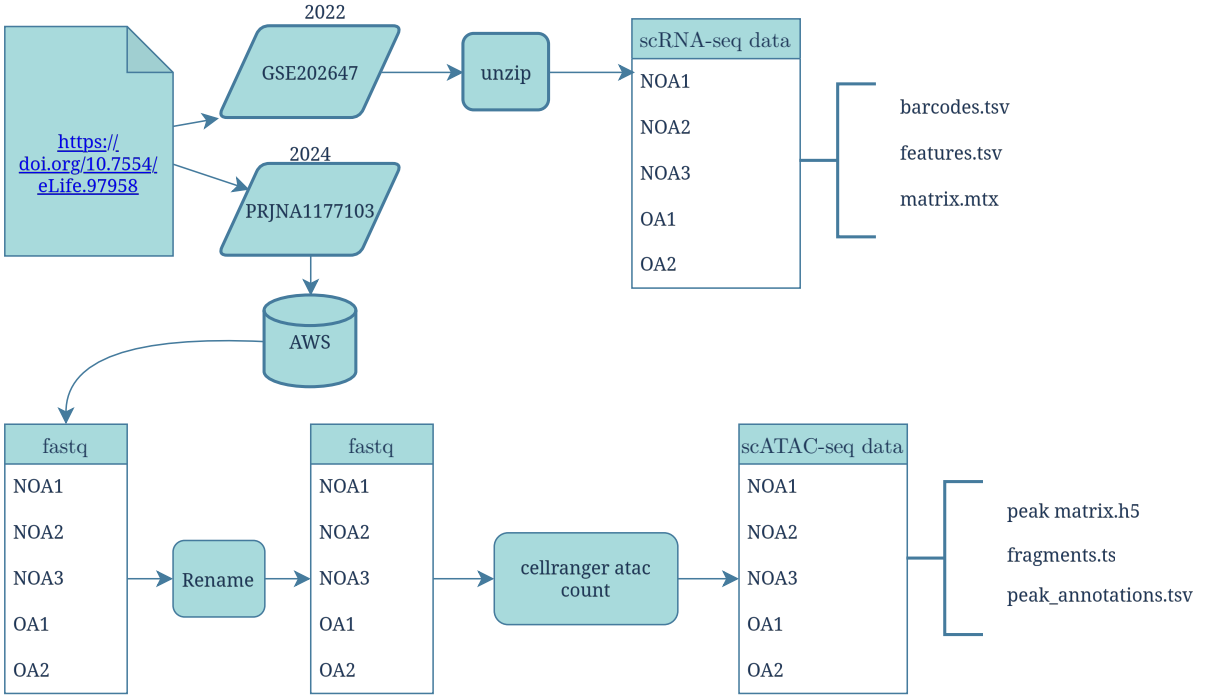

Figure 1: Stage 1 schematic of data aquisition and preparation.

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 [1] was used. This environment includes the scverse's anndata [2], mudata [3], scanpy [4], muon [5] and scvi-tools [6] packages, as well as full PyTorch [7] 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+ [8] suite.

2 Stage 1: Data aquisition and preparation

- For the testis dataset we could not find a cell matched RNA+ATAC -> seperate RNA and ATAC from same donors.
- Match by celltype instead of per cell.

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

[2] 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

[3] 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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

Appendix A Cellranger

Appendix B scRNA-seq

Appendix C scATAC-seq

Appendix D pycistopic workflow