Imperial College London
Institute of Clinical Sciences

# Dynamics of the Gene Regulatory Landscape in Human Spermatogenesis

Jahnavi Bhaskaran

August 2023

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy of Imperial College London

# Declaration

I certify that the work presented in this thesis is my own, and that any work carried out by others has been properly acknowledged and appropriately referenced in the text.

London, 14 August 2023

Jahnavi Bhaskaran

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

# Abstract

Spermatogenesis is a complex process in which cells undergo extensive changes, including the division of their genetic material and the exchange of histones with protamines, to ultimately produce mature haploid sperm. Little is known about the gene regulatory mechanisms driving spermatogenesis. Errors in this process can lead to infertility, which affects 10-15% of couples worldwide. Approximately half of these are cases of male factor infertility (Organization et al., 2023; Agarwal et al., 2021). However, only $\sim$30% of male infertility cases have a known cause (Tüttelmann et al., 2018).

In this thesis, we aim to better understand the process of human spermatogenesis. For this, we have used testis biopsies of men with healthy sperm production to generate single-nucleus multiome data capturing gene expression and chromatin accessibility measurements from the same cell. We have obtained more than 19,000 high quality nuclei from three patients, representing all the expected germline stages as well as supporting somatic cell types. By providing a direct association between the two captured modalities, multiome analysis reveals novel regulators of lineage-specific genes in addition to identifying rare cell populations. Using this technique, I have identified eight new subpopulations of spermatogonia, including an intermediate state whose gene expression resembles undifferentiated spermatogonia but whose transcription factor motif accessibility resembles differentiating spermatogonia. I have also captured two separate branches of undifferentiated spermatogonia during commitment to differentiation. Each branch displays a distinct pattern of niche communication, including signalling interactions that are misregulated in infertility. Additionally, I have utilized the multiome data to computationally separate haploid cells into X- and Y-spermatids. Following this, I have identified Y-spermatid-specific enrichment of ZNF28, a DNA-binding zinc-finger protein that is hypothesized to bind and silence SVA retrotransposons in the genome (Fukuda, Makino, et al., 2022). Lastly, we have established a novel sorting strategy to enrich for the key germ cell stages during human spermatogenesis. This has allowed the generation of high-resolution 3D genome maps of these stages, which capture dynamic 3D chromatin organization across sperm development, including the formation of the XY body during meiosis. Taken together, this thesis elucidates the dynamic gene regulatory landscape driving human spermatogenesis, which will further be highly beneficial for the study of male infertility.

To Amma and Appa

*To the eternally curious*

# Acknowledgements

I have had the privilege of meeting some incredible people who, in their unique ways, have been great sources of inspiration and support to me.

Firstly, I would like to thank my group leader Juanma Vaquerizas, for enthusiastically taking on a student who had little experience in computational analysis but a great desire to explore the field. Juanma's scientific curiosity is very infectious and this is evident in the exciting research pursued by the lab. I would also like to thank him and all members of the lab for creating an inspiring research environment. I will always be grateful for the support and feedback I have received from all of you. A special thanks to Liz Ing-Simmons, who formed the other half of the sperm duo. Every one of our interactions has been valuable to me and I have learnt so much from you, both science-related and otherwise. I would also like to say a big thank you to our collaborators at the Centrum für Reproduktionsmedizin und Andrologie (CeRA), Sara Di Persio, Nina Neuhaus, Sandra Laurentino and Lara Marie Siebert-Kuss, the hardcore sperm experts whose inputs have been crucial to the progress of the project. I have always been in awe of your depth of knowledge and your efficiency. The project also greatly benefited from the works of Noura Maziak, Virginie Tissières and Paul-Georg Majev. Thank you for your collaborative spirit and readiness to share your knowledge with me. Also, a big hug to Liz and Maria Rigau, who proofread the thesis in record time, and provided invaluable feedback.

I would like to acknowledge Fabian Groll, the in-house Snakemake expert for making everyone's analysis life easier. I greatly enjoyed my interactions with the ever enthusiastic Srishti Arya and her bottomless pile of cat photos. I would like to extend my appreciation to Sara Llorente Armijo, whose pursuit of dance and climbing apart from science, has been a great source of inspiration to me. I am also grateful for my interactions with Benjamín Hernández-Rodríguez and Quirze Rovira who always knew every work of research that exists on the 3D genome and transposable elements. I am happy to have met Irina Balaguer, Christos Kalaitzis and Melanie Chan, the future drivers of sperm science and multiomics in the lab.

I am grateful to my assessors Dr. James Turner, Dr. Harry Leitch, Dr. Malte Spielmann and Prof. Dr. Ralf Adams for their valuable feedback in driving the project forward. I would like to extend a special thanks to Prof. Petra Hajkova, whose encouragement and reassurance couldn't have come at a better time. I am indebted to Martin Wild, Claire Marsh and Poppy Brown, the PhD coordinators at the Max Planck Institute for Molecular Biomedicine (MPI) and the MRC London Institute of Medical Sciences (LMS). Thank you for everything you do behind the scenes to ensure a good experience for the PhD students. I am also grateful for the funding from the MPI

and Imperial College London that has supported my PhD.

During the course of my PhD, I have had the lovely opportunity to experience two different countries and cultures, which gave me the chance to meet some remarkable people.

Für meine deutsche Familie. Amaals, 넌 내가 뮌스터에서 제일 좋아하는 것 중 하나야 (네 음악 취향은 의심스럽지만). 내가 너에게 고마운 것들을 다 나열하면 논문 한 편은 써야 할거야. 그냥 말할게, 내 친구가 되어줘서 고마워. I am also happy to have met the lovely Pongal Potluckers, Nikita Chellam, Neeraja, Lakshmi and, pro-hiker Priya. Sara Cruz Molina, realmente hiciste de Münster un hogar para mí. Gracias por tu inmenso cariño, por la mejor tarta de zanahoria del mundo y por hacerme dar el paso para empezar a correr. I have found friendship in the most unexpected places. Susanne Hefti, ich werde für immer beeindruckt von und dankbar für die Offenheit sein, mit der Sie mich in Ihrem Leben und Zuhause willkommen geheissen haben. Ich bin stolz darauf, eine entscheidende Rolle bei der Entdeckung Ihrer Liebe zu Ghee gespielt zu haben.

I am forever indebted to Nina Schmolka, who has always had my best interests at heart. Thank you for being my mentor, my friend and my voice of reason.

I'd like to say a big thanks to The Bhavan for nurturing my interests outside science. A musical note, a beat of the Mridangam or the sound of the Salangai escapes through every door of the building. My time at the Bhavan has been very enriching and there hasn't been one weekend in the last 2 years that I have not spent there. The perpetual flow of memes with my Boyle cousins back home, Sam, Anu, Teju and Shruti has also been a staple and vital part of my day. Back in London, I would like to extend my appreciation to my lovely flatmates Margot, for panicking with me about the thesis deadlines, and Kiara, for not panicking with me about the thesis deadlines.

London, thank you for being everything I'd hoped for and Münster, thank you for being so much more. Kumudavalli, you have been my ride of freedom, literally, and I enjoyed every experience that much more when I went with you.

Amma. Appa. Everything I am is because of you (and I am not just referring to our shared genes). Thank you for being my biggest supporters and my biggest critics. Jagan, thank you for your constant encouragement, for being my advocate, and for wanting this for me as much as I have wanted it for myself, if not more.

Lastly, Janu, thank you for having my back. Nee gethu.

# Contents


Abstract . . . . . 4
Dedication . . . . . 5
Acknowledgements . . . . . 7
Table of Contents . . . . . 8
List of Figures . . . . . 15

**1 Introduction** **15**
1.1 Sex, Chromosomes and Rock 'n Roll . . . . . 15
1.2 Spermatogenesis in the adult testis . . . . . 15
1.2.1 Germ cell development . . . . . 16
1.2.1.1 Spermatocytogenesis . . . . . 16
1.2.1.2 Spermatidogenesis . . . . . 20
1.2.1.3 Spermiogenesis . . . . . 20
1.2.1.4 Spermiation . . . . . 22
1.2.2 Somatic cells in the testis . . . . . 22
1.3 Characterization of spermatogenesis . . . . . 23
1.3.1 Spermatogenesis through the lens . . . . . 23
1.3.1.1 A historical perspective . . . . . 23
1.3.1.2 Current trends . . . . . 23
1.3.2 Transcriptomic characterization . . . . . 26
1.3.2.1 single cell/nuclear RNA-seq technology . . . . . 26
1.3.2.2 Key findings from single cell/nuclear transcriptomic studies of spermatogenesis . . . . . 27
1.4 Gene regulation during human spermatogenesis . . . . . 32
1.4.1 Chromatin accessibility . . . . . 32
1.4.1.1 snATAC-seq/multiome snATAC-seq + snRNA-seq technology . . . . . 32
1.4.1.2 Key findings from snATAC-seq of human spermatogenesis . . . . . 32
1.4.2 Hierarchical organization of the 3D genome . . . . . 33
1.4.2.1 *in situ* Hi-C . . . . . 33
1.4.2.2 Key findings from 3D genome analysis of spermatogenesis . . . . . 35
1.5 Aims of the thesis . . . . . 39

**2 Methods** **40**
2.1 Selection of human testicular biopsies . . . . . 40
2.2 Preparation of single cell suspensions from human adult testis biopsy . 40

2.3 Generation of paired snATAC-seq and snRNA-seq libraries . . . . . 41
2.4 snATAC-seq and snRNA-seq analysis . . . . . 41
2.4.1 Read mapping and quantification . . . . . 41
2.4.2 RNA library QC . . . . . 42
2.4.3 ATAC library QC . . . . . 42
2.4.4 Peak calling . . . . . 42
2.4.5 Normalization, sample integration, dimensional reduction . . . . 42
2.4.5.1 RNA-seq assay . . . . . 42
2.4.5.2 ATAC-seq assay . . . . . 43
2.4.6 Weighted nearest neighbour (WNN) identification, clustering, and cluster annotation . . . . . 43
2.5 scRNA-seq analysis for scRNA vs snRNA comparison . . . . . 43
2.5.1 Read mapping . . . . . 43
2.5.2 Quality control filtering . . . . . 43
2.5.3 Normalization, sample integration, dimensional reduction . . . . 44
2.5.4 Clustering and cluster annotation . . . . . 44
2.6 Subclustering of spermatogonia clusters . . . . . 44
2.6.1 Normalization, sample integration, dimensional reduction . . . . 44
2.6.1.1 RNA-seq assay . . . . . 44
2.6.1.2 ATAC-seq assay . . . . . 44
2.6.2 Weighted nearest neighbour (WNN) identification, clustering, and cluster annotation . . . . . 44
2.7 Differential gene expression . . . . . 45
2.8 GO term enrichment . . . . . 45
2.9 Differential accessibility . . . . . 45
2.10 Motif activity . . . . . 46
2.11 Differential motif activity . . . . . 46
2.12 Correlation analysis . . . . . 46
2.13 TF footprinting . . . . . 47
2.14 Cell cycle analysis . . . . . 47
2.15 Slingshot trajectory inference . . . . . 47
2.16 MultiVelo analysis . . . . . 48
2.17 Cell-cell communication analysis . . . . . 48
2.17.1 Predicting Ligand-Receptor Interactions . . . . . 48
2.18 Conjugation of fluorophore to antibodies . . . . . 49
2.19 Whole mount staining of human seminiferous tubules . . . . . 50
2.20 FACS staining of testicular single cell suspension . . . . . 50
2.21 Low input Micro-C . . . . . 51
2.21.1 Crosslinking . . . . . 51
2.21.2 Preparation of Concanavalin A-coated beads . . . . . 51
2.21.3 Chromatin digestion by MNase . . . . . 51
2.21.4 End repairing and labeling . . . . . 52
2.21.5 Proximity ligation and removal of biotin-dNTP from un-ligated ends 52
2.21.6 DNA purification and Library Preparation . . . . . 52
2.22 Micro-C analysis . . . . . 53

**3 Generation and characterization of joint snATAC-seq + snRNA-seq multiome datasets of the adult human testis 54**

3.1 Introduction . . . 54
3.2 Optimization of nuclei isolation from the adult human testis . . . 55
3.3 snATAC-seq + snRNA-seq multiome captures all known cell types of the adult human testis . . . 60
3.4 snRNA-seq captures more spermatogonia and fewer spermatids than scRNA-seq . . . 64
3.5 Discussion . . . 68

**4 Dynamics of spermatogonia homeostasis in the adult human testis 72**
4.1 Introduction . . . 72
4.2 snATAC-seq + snRNA-seq multiome identifies new subclusters of spermatogonia in the adult human testis . . . 73
4.3 Gene expression profile of Early State 2 resembles undifferentiated spermatogonia . . . 75
4.4 Transcription factor motif activity profile of Early State 2 resembles differentiating spermatogonia . . . 77
4.5 State 0B exhibits a higher enrichment of mitochondrial transcripts and proliferative cells compared to State 0A . . . 80
4.6 Trajectory analysis of the spermatogonial compartment is inconclusive 82
4.7 State 0A and State 0B exhibit differences in niche signalling . . . 88
4.8 Discussion . . . 93

**5 Comparison of chromosome X- and Y-bearing spermatids 99**
5.1 Introduction . . . 99
5.1.1 Sex chromosomes in spermiogenesis . . . 99
5.1.2 Putative differences between chromosome X- and Y-bearing spermatids . . . 100
5.2 snATAC-seq can be used to classify spermatids into chromosome X- and chromosome Y-bearing . . . 101
5.3 *ZNF28* is enriched in Early Stage Y-spermatids . . . 105
5.4 X- and Y-spermatids show differential accessibility at gene promoters . 105
5.5 ZNF:C2H2-ZF and NHLH/TFAP:bHLH motifs show higher accessibility in Late Y-spermatids . . . 109
5.6 Discussion . . . 110

**6 Dynamics of 3D genome organization across adult human spermatogenesis 115**
6.1 Introduction . . . 115
6.2 Establishment of novel 5-colour FACS strategy to enrich for key developmental stages of human spermatogenesis . . . 116
6.3 Generation of Micro-C maps of 3D genome changes across adult human spermatogenesis . . . 123
6.4 CTCF footprint decreases during adult human spermatogenesis . . . 127
6.5 Discussion . . . 128

**7 Summary and conclusions 132**
7.1 Joint profiling of gene expression and chromatin accessibility in the adult human testis at the single cell level . . . 132

7.2 Characterization of the human spermatogonial compartment and its interaction with the niche . . . 133
7.3 Genomic and transcriptomic differences between X- and Y-spermatids 134
7.4 The dynamic 3D genome organization across adult human spermatogenesis . . . 134
7.5 Future perspective . . . 135
7.5.1 Spermatogonial homeostasis . . . 135
7.5.2 Escape from post-meiotic persistence of MSCI . . . 136
7.5.3 Post-transcriptional regulation of spermatogenesis . . . 136
7.5.4 Dynamics of expression of transposable elements in adult human spermatogenesis . . . 137
7.5.5 3D genome organization across adult human spermatogenesis 137

**Bibliography** **139**

**A Clinical parameters of samples processed in the multiome study** **155**

**B Antibody Dilutions** **157**

**C Journal Article (submitted)** **159**

# List of Figures


1.1 Ploidy vs number of chromatids. . . . . 16
1.2 Schematic of meiosis. . . . . 17
1.3 Sequential stages of Prophase I. . . . . 18
1.4 The synaptonemal complex. . . . . 18
1.5 The XY body. . . . . 19
1.6 Spermiogenesis. . . . . 21
1.7 Sertoli cells. . . . . 23
1.8 Light microscopy characterization of the human spermatogonia. . . . . 24
1.9 The seminiferous tubule cycle in the human testis. . . . . 25
1.10 Droplet-based sc/snRNA-seq technology from 10x Genomics. . . . . 26
1.11 Spatial transcriptomics with Slide-seq . . . . . 27
1.12 The human testis from the fetus to the adult. . . . . 29
1.13 Hierarchical organization of the 3D genome. . . . . 34
1.14 *in situ* Hi-C protocol and computational processing of the output. . . . 35
1.15 Representation of the hierarchical organization of the genome in the Hi-C heatmap. . . . . 36

3.1 Optimization of lysis buffer composition for nuclei isolation from adult human testicular cells. . . . . 56
3.2 Optimization of lysis time for nuclei isolation from HEK-293T cells. . . . 57
3.3 Optimization of lysis time for nuclei isolation from HEK-293T cells and adult human testicular cells. . . . . 58
3.4 Optimization of lysis time for nuclei isolation from adult human testicular cells. . . . . 59
3.5 10x snATAC + snRNA multiomic library generation from MeOH-fixed adult human testicular cells. . . . . 60
3.6 snATAC-seq + snRNA-seq multiomic characterization of the adult human testis. . . . . 61
3.7 Heatmap of gene expression across adult human spermatogenesis. . . 63
3.8 Examples of dynamic chromatin accessibility at selected marker genes. 64
3.9 Comparison of cell proportions between snRNA-seq and scRNA-seq of the adult human testis. . . . . 65
3.10 Comparison of gene expression between snRNA-seq and scRNA-seq of the adult human testis. . . . . 67
3.11 Number of differentially expressed genes per cluster between snRNA-seq and scRNA-seq of the adult human testis. . . . . 68
3.12 Gene Ontology enrichment analysis of differentially expressed genes between snRNA-seq and scRNA-seq of the adult human testis. . . . . 69

4.1 snATAC-seq + snRNA-seq multiomic characterization of spermatogonia subclusters in the adult human testis. . . . . 74
4.2 Marker gene expression across the spermatogonia subclusters in the adult human testis. . . . . 75
4.3 Gene expression characterization of the spermatogonia subclusters in the adult human testis. . . . . 76
4.4 Genome-wide differential chromatin accessibility characterization of the spermatogonia subclusters in the adult human testis. . . . . 78
4.5 Differential motif activity characterization of the spermatogonia subclusters in the adult human testis. . . . . 79
4.6 Transcription factor footprinting profiles across the spermatogonia subclusters in the adult human testis. . . . . 81
4.7 Exploration of confounding effects on the clustering of spermatogonia cell types in the adult human testis. . . . . 82
4.8 Cell cycle analysis with Tricycle. . . . . 83
4.9 Trajectory inference of the spermatogonia subclusters with Slingshot. . 84
4.10 Trajectory inference of the spermatogonia subclusters with MultiVelo. . 87
4.11 Number of ligand-receptor interactions in the adult human testis. . . . 89
4.12 Strength of ligand-receptor interactions in the adult human testis. . . . 90
4.13 Differential LRI between 8: State 0A and 5: State 0B (PIWIL4+). . . . . 91
4.14 Differential LRI between 3: State 0A and 2: State 0B (PIWIL4-). . . . . 92

5.1 The human pseudoautosomal region (PAR). . . . . 101
5.2 Distribution of reads from chromosome X or Y in human testicular germ cells. . . . . 102
5.3 Classification of spermatids into chromosome X- or Y-bearing. . . . . 103
5.4 Fraction of transcripts or snATAC-seq fragments in peaks from the X or Y chromosome in cells classified as X- or Y-spermatids using method employed in Murat et al., 2023 or this study. . . . . 104
5.5 Differential gene expression between X- and Y-spermatids. . . . . 106
5.6 Genomic distribution of differentially accessible peaks between X- and Y-spermatids. . . . . 107
5.7 Genomic loci of X- or Y-spermatid-specific accessibility. . . . . 108
5.8 Accessibility of snATAC-seq fragments in peaks in X- and Y-spermatids categorized based on overlap with significantly enriched genes. . . . . 109
5.9 Expression of genes in X- and Y-spermatids categorized based on overlap with significantly accessible peaks. . . . . 110
5.10 Differential motif accessibility between X- and Y-spermatids. . . . . 111
5.11 Footprint and gene expression of TFs predicted to bind to Y-spermatid-specific accessible motifs. . . . . 112

6.1 FACS strategy to enrich for major germ cell types in the adult human testis. . . . . 117
6.2 Testing abcam-mMAGEA4-Dy550 antibody for pan-spermatogonia cytoplasmic marker MAGEA4. . . . . 118
6.3 Testing mUTF1-Dy488 antibody for undifferentiated spermatogonia nuclear marker UTF1. . . . . 119
6.4 Testing BL-mKIT-Al647 and BD-mKIT-APC antibody for differentiating spermatogonia surface marker KIT. . . . . 120

6.5 Testing gKIT-Dy550 and mDMRT1-Al647 antibody for differentiating spermatogonia surface marker KIT and nuclear marker DMRT1. . . . . 121
6.6 FACS gating sequence to enrich for key germ cells from the adult human testis. . . . . 122
6.7 Cell counts of FACS-sorted subpopulations of key germ cells from the adult human testis. . . . . 123
6.8 Quality control of Micro-C libraries across adult human spermatogenesis.124
6.9 Micro-C contact frequency heatmaps across adult human spermatogenesis. . . . . 126
6.10 CTCF/CTCFL footprint in the adult human testicular germ cells. . . . . 128

# Chapter 1

# Introduction

*"...and I have seen so great a number of living animalcules in it, that sometimes more than a thousand were moving about in an amount of material the size of a grain of sand."*

- Antonie van Leeuwenhoek (November 1677), Letter 35 [22]; Addressed to: W. Brouncker, President of the Royal Society

## 1.1 Sex, Chromosomes and Rock 'n Roll

Sexual reproduction involves the union of genetic material from two individuals of different sexes. To maintain the number of chromosomes in the sexually reproducing species, the cells that fuse together to form the new organism must have only half the genetic material. Within a cell, the genetic material is organized into thread-like structures called the chromosomes, and before DNA replication, each chromosome exists as a single copy called a chromatid. The number of chromatids per chromosome represents the DNA content of a cell (C), while the number of the complete set of chromosomes defines the "ploidy" of a cell (N) (Britannica, 2020) (Figure 1.1). Most mammals, including humans, carry one set of chromosomes from each of the two parents. Thus, they are diploid (2N) organisms and, after mitosis, contain a total DNA content of 2C. Each chromosome from the maternal set resembles a chromosome from the paternal set in terms of its length, centromere position, and the set and order of genes contained. Such pairs of chromosomes are called homologous chromosomes.

## 1.2 Spermatogenesis in the adult testis

Gametogenesis describes the generation of haploid (1N) gametes from diploid (2N) cells, and spermatogenesis describes the production of the gametes in males.

In mammals and other amniotes, spermatogenesis primarily occurs within the seminiferous tubules of the testis. Here, stem cells called spermatogonia reside on the

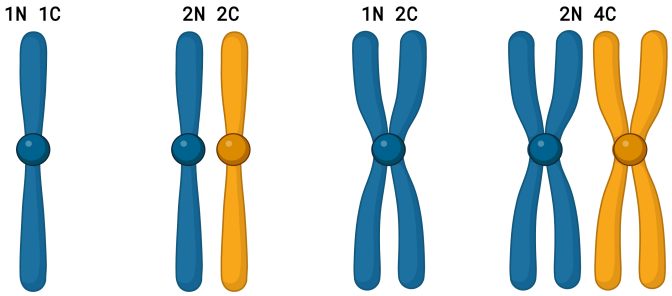



Figure 1.1: ***Ploidy vs number of chromatids.*** Maternal and paternal DNA is represented by different colours.

periphery of the tubule. A subset of the undifferentiated spermatogonia commit to differentiation to become the differentiating spermatogonia. These cells duplicate their DNA to form primary spermatocytes, which undergo meiosis to produce secondary spermatocytes and subsequently, round haploid spermatids. The round spermatids then undergo extensive differentiation to form elongated spermatids. As cells advance through spermatogenesis, they progressively move closer to the central lumen of the tubule and the elongated spermatids are finally released into the lumen. Upon release, the spermatids undergo the last steps of maturation in the epididymis, where they are stored until ejaculation. This process of germ cell development is also supported by a number of somatic cells that reside in the testis (Gilbert and Barresi, 2000).

In humans, the entire process of spermatogenesis takes around 74 days (Heller, 1964) and produces 35-40 million spermatozoa per ejaculate (Boitrelle et al., 2021).

## 1.2.1 Germ cell development

Spermatogenesis can be divided into the following sequential stages:

### 1.2.1.1 Spermatocytogenesis

The production of secondary spermatocytes from undifferentiated spermatogonia is called spermatocytogenesis. In the human testis, two types of undifferentiated spermatogonia have been identified based on nuclear staining and morphology - Type A dark ($A_{dark}$) and Type A pale ($A_{pale}$). Based on the expression of the proliferation marker MCM7 (Padmanabhan et al., 2004; Choy et al., 2016) in immunofluorescence staining of human seminiferous tubules (Caldeira-Brant et al., 2020) and on the incorporation of 3H-thymidine in monkey testes (Clermont, 1969), $A_{dark}$ is considered to represent a quiescent population of reserve stem cells and $A_{pale}$ is believed to represent the active stem cell population of the testis (Clermont, 1972; Van Alphen et al., 1988). An $A_{dark}$ spermatogonium (2N, 2C) has been proposed to undergo asymmetric division to produce another $A_{dark}$ spermatogonium and an $A_{pale}$ spermatogonium (2N, 2C) while the $A_{pale}$ spermatogonium undergoes mitosis to form a clone of cells (Clermont, 1969; Clermont, 1972). Due to incomplete cytokinesis, $A_{pale}$ cells remain connected by cytoplasmic bridges (Dym and Fawcett, 1971). The linked cells then differentiate to form Type B differentiating spermatogonia (2N, 2C). Type B spermatogonia undergo further mitosis to produce primary spermatocytes. These cells replicate

their DNA, such that each chromosome now has two identical sister chromatids (2N, 4C) that are held together by cohesin (Peters and Nishiyama, 2012). Following DNA replication, the primary spermatocytes enter meiosis.

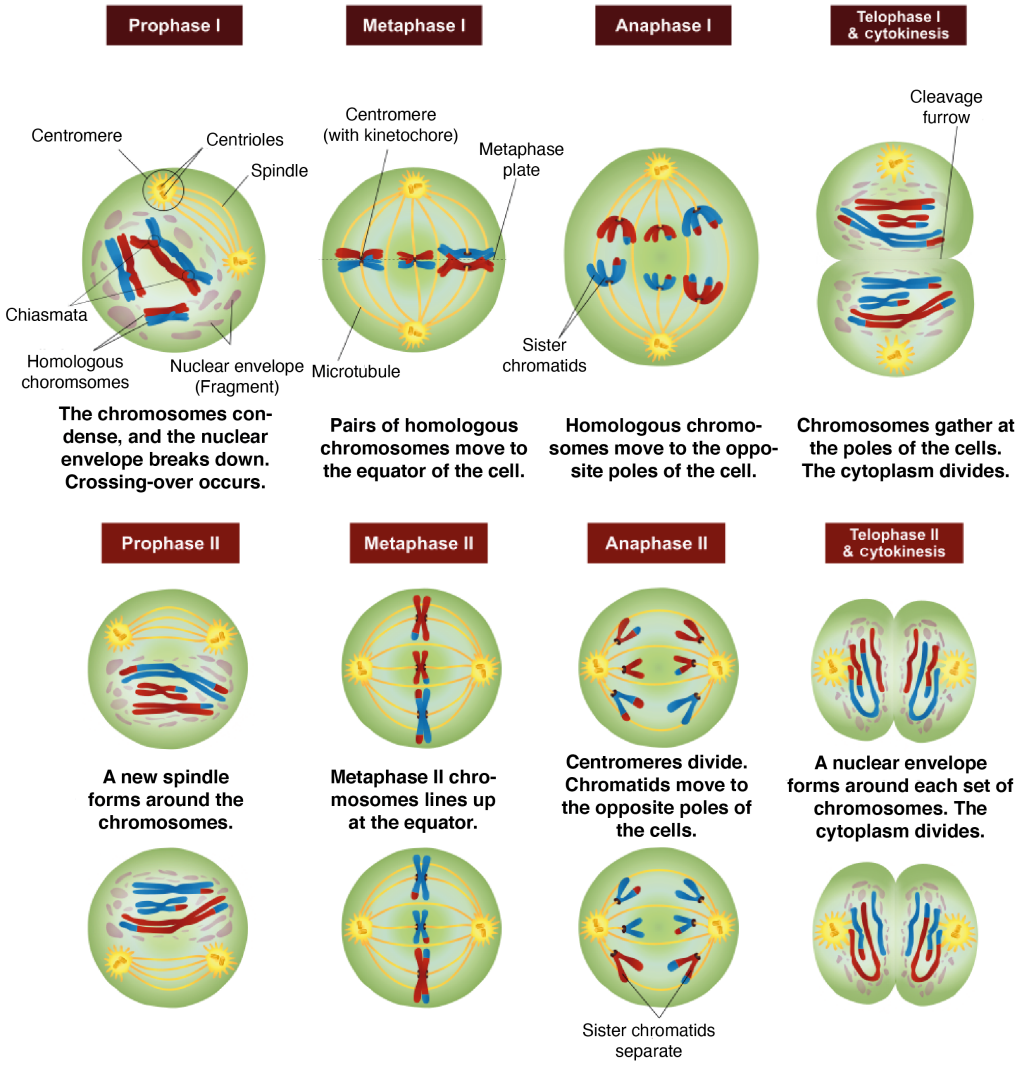



Figure 1.2: ***Schematic of meiosis.*** Meiosis results in the reduction of the number of chromosomes within a cell to produce haploid gametes. (Adapted from illustration by Ali Zifan using information from Reece et al., 2015)

Meiosis is a type of cell division process that produces haploid cells, and it can be broken down into meiosis I and meiosis II. Meiosis I is called "reductional division" because it separates the pair of homologous chromosomes to produce haploid secondary spermatocytes. The stages of meiosis I include:

**Prophase I** Prophase I is the longest phase of meiosis. In prophase I, the maternal and paternal homologous chromosomes pair up and exchange genetic material in a process called homologous recombination (Figure 1.3). Such an exchange facilitates genetic variation in sexually reproducing organisms. prophase I includes:

**Leptotene** In leptotene, the genome is transcriptionally inert (Monesi, 1964; Paronetto et al., 2011) and the chromosomes start becoming visible as individual strands in the

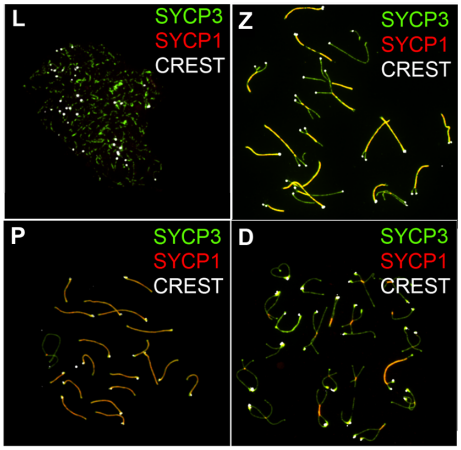



Figure 1.3: ***Sequential stages of Prophase I (Qiao et al., 2012).*** Surface spread of mouse spermatocytes at leptotene (L), late zygotene (Z), pachytene (P) and diplotene (D) stained for CREST that binds to centromere (Earnshaw and Rothfield, 1985), SYCP3 (Bollschweiler et al., 2019) and SYCP1 (Schücker et al., 2015), the lateral and central element protein of the synaptonemal complex.

nucleus. SPO11, a meiosis-specific endonuclease, creates double strand breaks on the DNA and initiates the process of recombination (Keeney, 2007). The DNA repair protein RAD51 and the meiotic recombinase protein DMC1 coat the resulting single strands of the DNA and invade the homologous chromosomes to achieve their pairing and co-alignment (Hinch et al., 2020). Leptotene is also characterized by the presence of histone H3 lysine dimethylation (H3K9me2) which, along with DNA methylation, is involved in the silencing of LINE1 transposable elements at this stage (Di Giacomo et al., 2013). Leptotene spermatocytes cross the blood-testis barrier to enter the adluminal compartment of the seminiferous tubule.

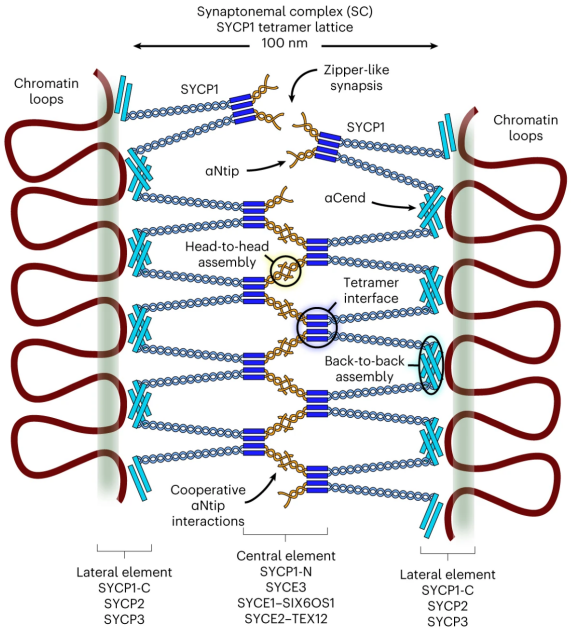



Figure 1.4: ***The synaptonemal complex (Crichton et al., 2023).***

**Zygotene** In the following stage of zygotene, the pairing of the homologous chromosomes is further stabilized through a zipper-like process of synapsis facilitated by the synaptonemal complex. The synaptonemal complex is a tripartite structure

that includes two lateral (one for each chromosome in the homologous pair) and one central element (Figure 1.4). It is assembled along the length of the chromosomes at the interface between homologous chromosome pairs to result in "bivalent" or "tetrad" chromosomes. During this process, the genome continues to remain transcriptionally silent (Monesi, 1964; Paronetto et al., 2011). From late zygotene onwards, H3K9me2 decreases and the silencing of LINE1 is achieved through post-transcriptional cleavage by Piwi protein Mili, apart from DNA methylation (Di Giacomo et al., 2013).

**Pachytene** In pachytene, the double strand breaks generated in leptotene are repaired and this is accompanied by active transcription from mid-pachytene onwards (Monesi, 1964; Paronetto et al., 2011). A subset of the double strand breaks are repaired by forming a crossing over between non-sister arms of the homologous chromosomes and this results in the exchange of genetic material between them (Keeney, 2007). During this process, only the autosomes exhibit complete synapsis. The sex chromosomes only synapse at the chromosomal ends, where they share homology over regions called pseudoautosomal regions (PARs) (Ciccodicola et al., 2000). Incomplete synapsis of the sex chromosomes triggers a process called Meiotic Sex Chromosome Inactivation (MSCI). During MSCI, the unsynapsed sex chromosomes are recognized by members of the synaptonemal complex and DNA repair machinery. This recognition leads to histone H2AX-phosphorylation-dependent silencing. The silencing is further stabilized by other modifications such as histone H2A ubiquitination, histone H3 lysine methylation (H3K4me) and replacement of histone H3 variants H3.1/H3.2 with H3.3 (Turner, 2015). During this process, the sex chromosomes form a spatially segregated chromatin territory called the XY body (Handel, 2004) (Figure 1.5).

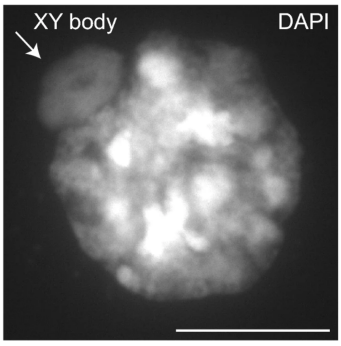



Figure 1.5: ***The XY body (Namekawa et al., 2006).*** DAPI counterstaining of mid-pachytene spermatocytes in the mouse germline prepared using a technique that preserves the 3D organization of the chromatin. The sex chromosomes form a spatially segregated chromatin territory called the XY body during Meiotic Sex Chromosome Inactivation (MSCI).

**Diplotene** Subsequently, the cells enter diplotene, during which the synaptonemal complex dissolves and the homologous chromosomes separate except at the regions of crossing-over called the chiasmata.

**Metaphase I** In metaphase I, the homologous pairs align along the metaphase plate. The homologous chromosomes are held along the plate due to the tension exerted by microtubules that connect the kinetochore of each homologous chromosome in the pair to the spindle poles on opposite sides of the cell. The establishment

of this tension is further aided by the cohesin complex holding the sister chromatids together, and by regions of crossing-over between non-sister chromatids (Lu, Zong, et al., 2012).

**Anaphase I** In anaphase I, cohesin along the chromosome arms is cleaved, and each chromosome in the homologous pair separates to opposite poles of the cell. However, cohesin around the centromere is protected by the Shugoshin protein from degradation (Yao and Dai, 2012). Hence, the sister chromatids remain together during this segregation process.

**Telophase I** In telophase I, the chromosomes have finished moving to the opposite sides of the cell, and meiosis I ends to produce secondary spermatocytes that each carry two sister chromatids of one of the homologous chromosomes (1N, 2C).

#### 1.2.1.2 Spermatidogenesis

Spermatidogenesis represents Meiosis II, when secondary spermatocytes rapidly divide to produce spermatids. Meiosis II is an "equational division" process and resembles mitosis. During this process, the sister chromatids are separated into different cells to produce haploid spermatids that carry only copy of the sister chromatids (1N, 1C) (Figure 1.2).

#### 1.2.1.3 Spermiogenesis

Haploid spermatids produced at the end of spermatidogenesis are small and circular cells. Spermiogenesis describes the subsequent stages of extensive differentiation that the round spermatids undergo to become the mature elongated spermatids. Key developments during this process include substantial changes in cell morphology and a tighter packaging of the DNA within the nucleus.

**Morphological changes** The morphological changes during spermiogenesis can be divided into four sequential stages:

**Golgi phase** During the Golgi phase, the round spermatids develop polarity and start forming the major parts of a mature spermatozoon: the head, the mid-piece and the tail. Simultaneously, the Golgi apparatus produces many lytic enzymes in preparation for the development of the sperm head at one end of the polarized cell. A single acrosomal vesicle is formed near the nuclear membrane, which fuses with Golgi-derived vesicles to form the acrosomal sac. Subsequently, the centrioles migrate to the other end of the cell.

**Cap phase** During the cap phase, the acrosomal sac develops into an enzyme-filled, membrane-bound compartment called the acrosome/acrosomal cap which is attached to the nuclear envelope on the anterior side of the cell. During fertilization, these enzymes are used to break down the outer membrane of the egg (Yanagimachi, 1994), and allow the joining of the sperm and egg nucleus. At the same time, the tail axoneme continuous growing on the posterior side of the cell. Within the axoneme, the microtubules are arranged in a “9+2” fashion, where a central pair

of microtubules is surrounded by 9 outer doublet microtubules. The outer doublets generate the forces to produce the waveform motion of the tail (Eddy, 2006).

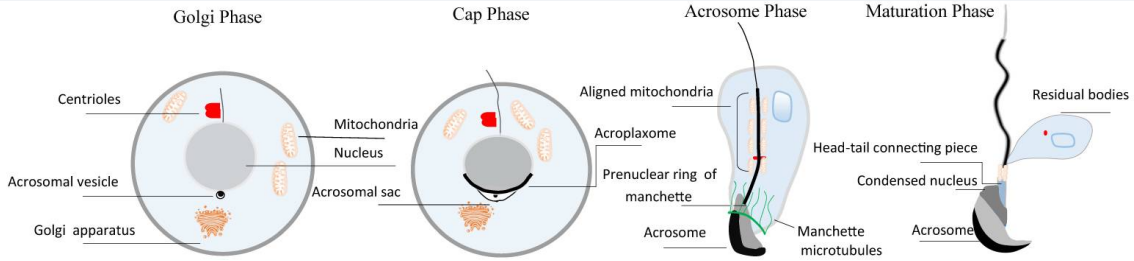



Figure 1.6: ***Spermiogenesis (Azhar et al., 2021).*** Early round spermatids undergo sequential differentiation to form the late elongated spermatids during spermiogenesis.

**Acrosome phase** During the acrosome phase, the acroplaxome complex which is attached to the growing acrosome cap starts descending on the posterior side towards the growing axoneme. This is accompanied by the development of a microtubule-containing structure called the manchette. Additionally, at the posterior end, the mitochondria gather together to form the mid-piece. The mitochondria provide the energy required for movement of the tail. These changes are accompanied by high condensation and transcriptional silencing of the nucleus as described in **Nuclear remodelling**. Furthermore, the developing germ cells orient themselves so that the growing tails are pointing towards the central lumen of the seminiferous tubule.

**Maturation phase** During the maturation phase, the cytoplasm remaining after nuclear condensation, called the residual body, is phagocytosed by a class of supporting somatic cells called Sertoli cells (Kerr and De Kretser, 1974).

**Nuclear remodelling** Spermiogenesis also includes considerable rearrangement of the DNA within the nucleus. During this process, the histones in the nucleus are sequentially replaced by specific transition proteins and protamines to allow tighter and more highly condensed packaging of the DNA.

The histone-to-protamine exchange begins with the eviction of nucleosomes through the incorporation of testicular and haplophase-specific histone variants (Boussouar et al., 2008; Rousseaux et al., 2008). This is accompanied by specific histone modifications such as H4 acetylation and H2A ubiquitination (Baarends et al., 1999; Grimes Jr and Henderson, 1984; Meistrich, Trostle-Weige, et al., 1992). The torsional stress encountered during the eviction of nucleosomes is released through the formation of DNA strand breaks (Marcon and Boissonneault, 2004). Subsequently, transition proteins (TPs), such as TNP1 and TNP2 in humans, are incorporated in to the DNA (Meistrich, Mohapatra, et al., 2003; Rathke et al., 2014; Bao and Bedford, 2016). Eventually, the TPs are replaced by a class of small, arginine-rich, nuclear proteins called protamines. Through their arginine-rich domain, protamines bind to the phosphate backbone of the DNA and fold it condensely into a toroid (Brewer et al., 1999). T. Wang et al., 2019 provide a detailed overview of the role of each of these modifications in spermiogenesis. While histones are replaced with protamines throughout the genome, $\sim$1% of mouse genome (Heijden et al., 2005) and 10-15% of human genome (Brykczynska et al., 2010; Gatewood et al., 1987) are still found to retain histones. These regions includes gene deserts (Carone et al., 2014; Yamaguchi et al.,

2018) and regions thought to be essential for spermiogenesis and early embryonic development (Hammoud et al., 2009).

#### 1.2.1.4 Spermiation

The final products of spermiogenesis are non-motile mature spermatozoa. During spermiation, the cells are released into the central lumen of the seminiferous tubule and transported to another convoluted tubule called the epididymis. Because spermatozoa are still non-motile, they are transported to the epididymis in the testicular fluid secreted by the Sertoli cells though peristaltic contractions of the tubule wall. Finally, the spermatids reach the epididymis, where they gain motility and are stored until ejaculation.

### 1.2.2 Somatic cells in the testis

Apart from the developing germ cells, the testis also contains a number of somatic cells that support germ cell development. The somatic cells are present both within the seminiferous tubule, such as the Sertoli cells, or in the interstitial space between the tubules, such as the Leydig cells and peritubular myoid cells (PTMs). Besides, the testis also contains blood vessels and immune cells.

**Sertoli cells** Sertoli cells are large tree-like cells that extend from the basal membrane to lumen of the seminiferous tubule. They are called "nurse" cells as they provide a variety of nutrients, hormones and growth factors to the developing germ cell at different stages of spermatogenesis (Mruk and Cheng, 2004). Their functions include the maintenance of the stem cell niche (Oatley et al., 2011), regulating entry into meiosis by influencing retinoic acid levels, a key signal for spermatogenesis (Raverdeau et al., 2012), and phagocytosis of the residual body during spermatid maturation (O'Donnell et al., 2011). Their structural functions include forming the blood-testis barrier to maintain the immune-privileged environment of the adluminal compartment of the tubule (França et al., 2016).

**Leydig cells** Leydig cells are the primary producers of testosterone, and they are located at the interstitial compartment between the seminiferous tubules (Zirkin and Papadopoulos, 2018).

**Peritubular myoid cells** Peritubular myoid cells (PTMs) are smooth muscle cells that surround the human seminiferous tubules in multiple layers and provide structural integrity to them. They are responsible for the peristaltic movements that aid in the transport of immotile spermatozoa from the lumen of the tubule towards the epididymis. In the adult human testis, two subgroups of adult PTMs have been identified, each possibly existing in a different layer around the tubule (Mayerhofer, 2013). One group is enriched for muscle contraction-related genes, and the other, termed fibrotic PTMs, is for ECM components-related genes (Di Persio, Tekath, et al., 2021). Located at the interface between the interstitial compartment and the Sertoli/germ cell compartment, PTMs exhibit a wide range of paracrine signalling, including those involved

in spermatogonia proliferation (Spinnler et al., 2010) and those in the deposition of extracellular matrix (Mayerhofer, 2013).

## 1.3 Characterization of spermatogenesis

### 1.3.1 Spermatogenesis through the lens

#### 1.3.1.1 A historical perspective

As with many biological discoveries of earlier times, anatomical descriptions and microscopy observations were the primary mode of cellular characterization of the spermatogenic process.

**The spermatozoa** In November 1677, for the first time, Johan Ham and Antonie van Leeuwenhoek observed spermatozoa in human semen under a light microscope. In a letter to the President of the Royal Society, van Leuwenhoek described seeing a vast number of "animalcules" that were smaller than red blood corpuscles, and contained a round body and a thin tail. He noted that the tail was transparent and as thin as one twenty-fifth of the body. He even recounted the forward motion of the spermatozoa and attributed it to the movement of their tails "like that of a snake or an eel swimming in water" (Howards, 1997; Leeuwenhoek and Rijnberk, 1939).



Figure 1.7: ***Sertoli cells. (From Sertoli E. Il Morgagni 1865). (Patriarca et al., 2019).*** The first sightings of the Sertoli cells.

**Somatic cells** This was followed by the discovery of the testicular interstitial Leydig cells by Franz Leydig, a German zoologist and a comparative anatomist, in 1850. He described them as a fatty "cell-like mass" in which the "seminiferous tubules are embedded" (Leydig, 1850; Christensen, 2007). In 1865, Italian histologist Enrico Sertoli first described the Sertoli cells, as being "irregularly cylindrical or conical" with branch-like extensions that also contain fine fat droplets (Sertoli, 1865; Patriarca et al., 2019).

#### 1.3.1.2 Current trends

Since then, histological evaluation of testes biopsies, including the characterization of cell morphology and the quantification of cell count, has become an integral part of

understanding healthy and altered spermatogenesis (McLachlan et al., 2007).

**Germ cell developmental stage** Early studies of fixed histological sections stained with hematoxylin and eosin have pioneered the detailed characterization of the germ cells at different stages of development (Clermont, 1963).

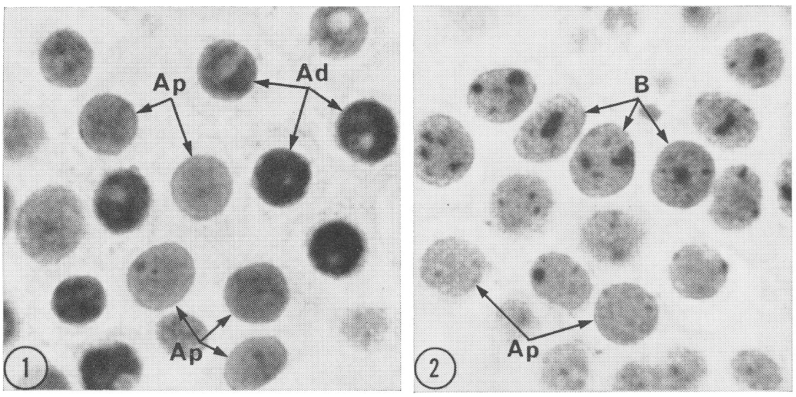



Figure 1.8: ***Light microscopy characterization of the human spermatogonia. (Clermont, 1966).*** The human spermatogonia subtypes were found to distinct nuclear staining.

Such characterizations have used nuclear staining to classify human spermatogonia into Types $A_{dark}$, $A_{pale}$, and B. $A_{dark}$ spermatogonia have a spherical, deeply stained nucleus with a large cavity in its centre. $A_{pale}$ spermatogonia contain a discoid, lightly-stained nucleus with no cavity in their nucleoplasm but show nucleolus attached to their nuclear membrane. Type B spermatogonia show a spherical, lightly-stained nucleus containing some intensely-stained granules.

Cells at the different stages of meiotic prophase I, including leptotene, zygotene and pachytene can also be distinguished based on the chromatin structure within the nucleus. Pre-leptotene spermatocytes are slightly smaller than B spermatogonia and contain deeply stained chromatin accumulated on the nuclear periphery and the nucleolus. Leptotene spermatocytes show their characteristic fine chromosomal filaments and pachytene spermatocytes form the biggest germ cells in the testis. Secondary spermatocytes have a spherical deeply-stained nucleus and are only slightly bigger than round spermatids.

Lastly, spermatids can be classified into different stages of spermiogenesis based on cellular and nuclear morphology. Early round spermatids have lightly-stained spherical nucleus (Sa) which starts to darken (Sb1) and elongate (Sb2). Then, the cytoplasm also starts to elongate with the flagellum becoming visible (Sc) and subsequently, the nucleus reduces in size (Sd1). Finally, the nucleus completely condenses and the germ cell becomes the free cell spermatozoa (Sd2) which is quickly released into the tubular lumen. Since germ cells move from the periphery towards the lumen during development, the location of the germ cell across the cross-section of the histological section further aids in the identification of the developmental stage.

**Staging of the tubule** The detailed description of the different germ cell stages by microscopy studies has revealed a "cycling" nature of the seminiferous tubule. Within the testis, spermatogonia residing in a small section of the seminiferous tubule collectively commit to differentiation at specific intervals. Based on radioactive labelling,

$A_{pale}$ spermatogonia in the human testis is estimated to enter active spermatogenesis at an interval of $\sim$ 16 days (Clermont, 1963). The subsequent progression of germ cell development to produce haploid spermatozoa also follows a rigid timescale. In humans, this process takes $\sim$ 74 days. Due to the simultaneous occurrence of spermatogonia commitment and spermatogenic progression at different periodicities, and the progressive movement of germ cells towards the tubule lumen during development, specific groups of cell stages, termed "cellular associations" can be observed in different sections of the seminiferous tubule. These cellular associations can be used to describe the cycle of the seminiferous tubule (Clermont, 1963; Amann, 2008). The organization of the human testis into six cellular associations is depicted in Figure 1.9.

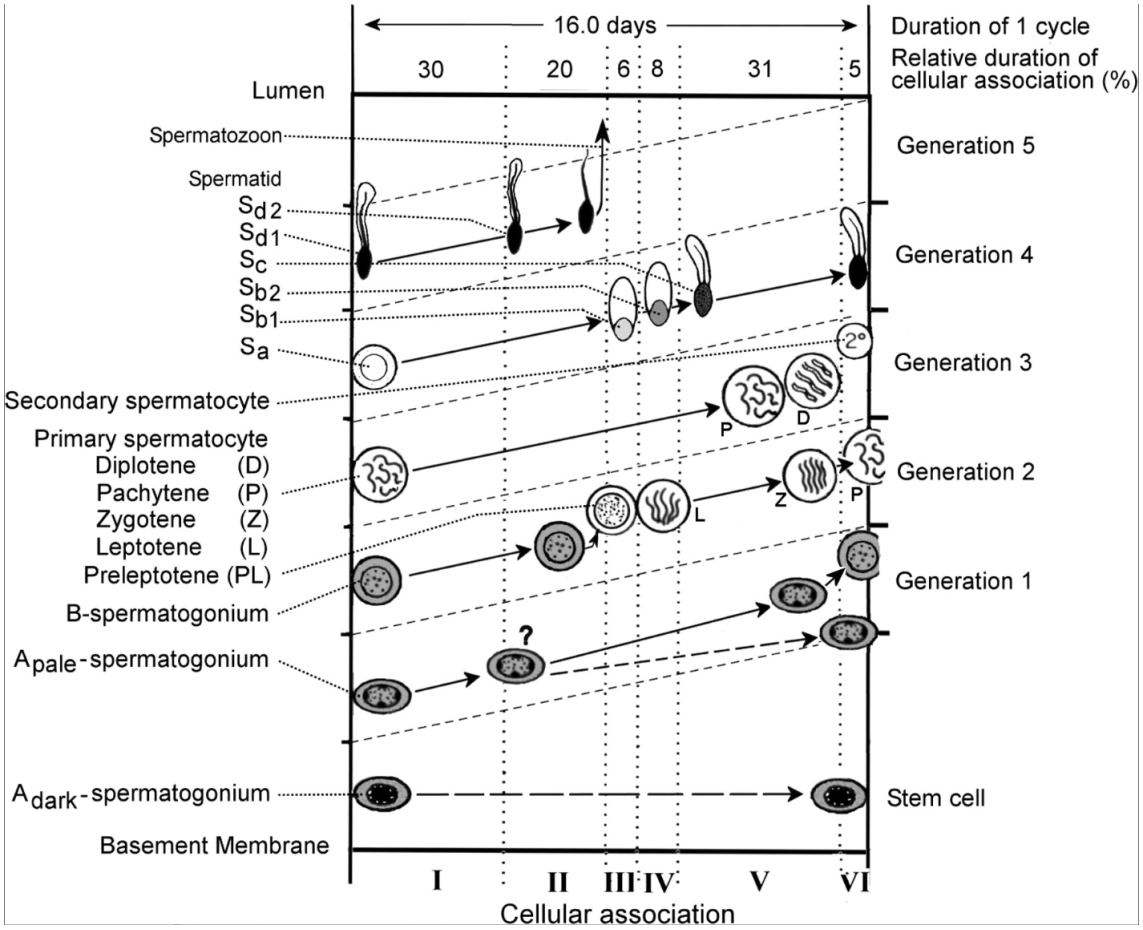



Figure 1.9: ***The seminiferous tubule cycle in the human testis (Clermont, 1963; Amann, 2008).*** The $A_{pale}$ spermatogonium in cellular association II has been marked with a "?" since it is still unclear if these are the first cells that commit to differentiation.

**Disease classification** Detailed description of cellular morphology and cell proportions within the testis also allows a descriptive classification of altered human spermatogenesis. This includes hypospermatogenesis, which indicates a reduction of the germ cells at different developmental stages to varying degrees, Sertoli cell-only syndrome which exhibits a complete absence of all germ cells, cell arrest at specific germ cell stages, etc (McLachlan et al., 2007).

### 1.3.2 Transcriptomic characterization

Over the past years, advances in technology has enabled researchers to capture the products of gene transcription and characterize a cell state based on the level and pattern of gene expression. This allows the discernment of changes within a cell that may not be visible under a microscope. Starting from a handful of genes, it is now possible to map global changes in gene expression, and at the level of single cells.

#### 1.3.2.1 single cell/nuclear RNA-seq technology

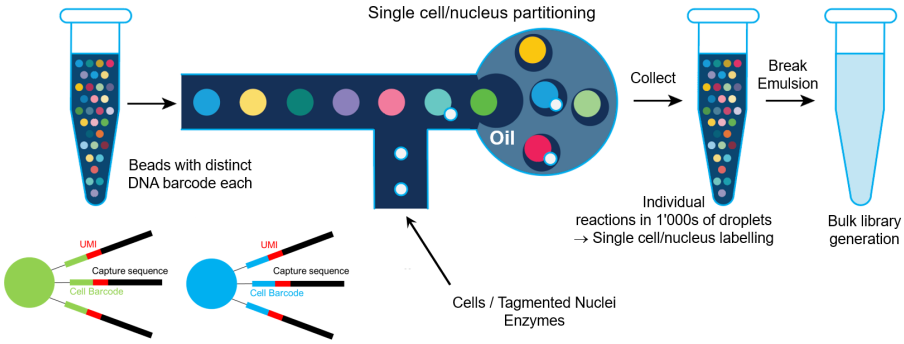



Figure 1.10: ***Droplet-based sc/snRNA-seq technology from 10x Genomics (*https://www.10xgenomics.com/ *n.d.;* https://wp.unil.ch/gtf/single-cell-experiments-10x/ *n.d.).***

**Droplet-based single cell generation** Droplet-based single cell partitioning technology employs microfluidic systems to compartmentalize cells into individual droplets for tagging. Platforms that employ droplet-based partitioning of cells include the Chromium Single Cell platform from 10x Genomics (*https://www.10xgenomics.com/* n.d.), and inDrop (Klein et al., 2015) and Drop-seq (Macosko et al., 2015) systems commercialized by Illumina (*https://www.illumina.com/* n.d.). In the 10x Chromium Single Cell platform, cells or nuclei in solution are partitioned with distinct gel beads into individual oil droplets (Figure 1.10). Each gel bead contains sequences called unique molecular identifiers (UMIs) that tag individual molecules. Additionally, each gel bead also contains a cell barcode, which tags all the transcripts arising from the same cell with the same barcode. Hence, the tagged transcripts from various cells can be pooled together for sequencing. The sequences from each transcript can be mapped back to the cell and the gene of origin. Thus, one can investigate changes across all the genes and across 10,000 individual cells in one experiment.

**Spatial transcriptomics** The droplet-based single cell library generation strategy requires dissociation of the tissue into single cell suspension. Thus, the spatial organization of the cells within the tissue is lost. Spatial transcriptomics addresses this challenge by using a cross-section of the tissue as starting material. In Slide-seq (Rodriques et al., 2019), one such spatial transcriptomics platform, gel beads containing cell barcodes and UMIs, similar to the ones used in Drop-seq, are coated as a monolayer on a rubber-coated glass coverslip. A thin layer of tissue section is then placed on the bead monolayer. The beads then capture and tag the RNA released from the tissue. As in droplet-based tagging, the RNA molecules from the coverslip

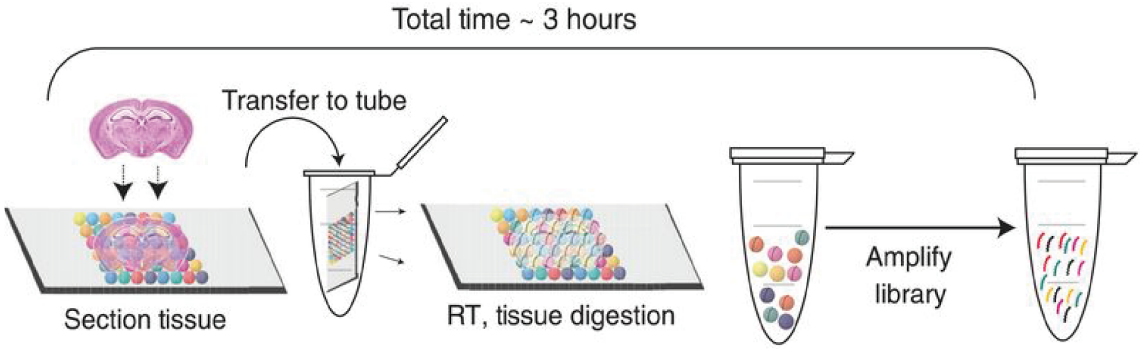



Figure 1.11: ***Spatial transcriptomics with Slide-seq (Rodriques et al., 2019).***

are then pooled together for sequencing. The cell types identified from the sequencing are computationally mapped back to the position along the cross-section of the slide. Thus, the technique captures the organization of the individual cells in the tissue, apart from their transcriptomes. Other spatial transcriptomics platforms include Visium by 10X Genomics (Ståhl et al., 2016), and GeoMx (Merritt et al., 2020) and CosMx (He et al., 2021) by Nanostring.

#### 1.3.2.2 Key findings from single cell/nuclear transcriptomic studies of spermatogenesis

Single cell (sc) and single nucleus (sn) transcriptomic approaches have allowed the characterization of the testicular tissue across multiple stages of development and disease. Key contributions to our understanding of the human testis from these approaches is summarized below.

**The adult human testis** Multiple studies have generated a transcriptional atlas of the adult human testis to capture the gene expression profiles that define the different germ and somatic cell types. In the UMAP representation of clusters arranged based on transcriptome similarity, the germ cell clusters form a continuous trajectory while the somatic cells form distinct individual clusters, replicating the biological arrangement of these cells in the testis. Undifferentiated spermatogonia express known SSC markers such as *UTF1* and *ID4*, differentiating markers express *KIT* and *DMRT1*, meiotic spermatocytes express *SYCP3* and *SPO11* and spermatids expresses protamine repackaging factors such as *TNP1* and *PRM2*. Apart from identifying known stage-specific markers, these studies have also discovered a plethora of new markers for the different testicular cells. They have also captured dynamic patterns of long non-coding RNAs across spermatogenesis, such as the enrichment of *XIST* in spermatogonia. Furthermore, repetitive elements are also found to show stage-specific expression, such as the enrichment of satellite elements in spermatid development. (Guo, Grow, Mlcochova, et al., 2018; Sohni et al., 2019; Shami et al., 2020; Hermann et al., 2018; Wang, Liu, et al., 2018; Xia et al., 2020). The transcriptomic profiles recapitulate the biological developmental trajectory of germ cells and capture patterns of transcriptional inactivation on sex chromosomes associated with Meiotic Sex Chromosome Inactivation (MSCI) (Guo, Grow, Mlcochova, et al., 2018; Wang, Liu, et al., 2018). Finally, snRNA-seq data has also been used to computationally distinguish X- and Y-spermatids and identify the relative enrichment of long intergenic non-coding

RNAs (lincRNAs) such as *FAM230C* and *FAM230F* in Y-spermatids (Murat et al., 2023).

**Germline-niche communication** Cell-cell communication has been inferred from sc/snRNA-seq data using prior knowledge of ligand-receptor pairs and their gene expression in pairs of cells in the transcriptome data. Overall, spermatogonia have more putative interactions with somatic cells than the other germ cells, in line with the blood-testis barrier preventing communication between meiotic and post-meiotic stages and the interstitial cells (Shami et al., 2020). Multiple studies also find ligands and receptors associated with WNT signalling in the testis. *WNT2B* is expressed in myoid cells and its receptors are expressed in primary spermatocytes, suggesting a possible role in meiosis (Shami et al., 2020). WNT activator *RSPO2* is also found in the germ cells while its receptor *LGR4* is found in PTMs (Sohni et al., 2019).

Spatial transcriptomics, by maintaining the arrangement of cells in the tissue, provides a more robust inference of cell-cell communication compared to sc/snRNA-seq techniques. It has identified three spatial clusters within the human spermatogonia compartment including one cluster formed by State 0, 0A, 0B and 1, another by State 2 and a third by State 3 with the cluster nomenclature as used in Di Persio, Tekath, et al., 2021 and this study. Moreover, the undifferentiated and differentiating spermatogonia were found to be differentially enriched for somatic cells in their microenvironment, including Leydig cells, endothelial cells, and macrophages, but not Sertoli cells and PTMs. The endothelial cells are proposed to regulate the stem cell niche through ephrin A1 (EFNA1)-ephrin type A receptor (EPHA) pairing with spermatogonia (Rajachandran et al., 2023; Chen, Murray, et al., 2021).

**Characterization of spermatogonia** In addition to capturing known biology, these studies have also identified new subtypes of testicular cells that are indistinguishable in light microscopy (Guo, Grow, Mlcochova, et al., 2018; Sohni et al., 2019). The transcriptomic profiles have enabled the classification of undifferentiated spermatogonia broadly into State 0, State 1 and differentiating spermatogonia into State 2 and State 3 (Guo, Grow, Mlcochova, et al., 2018; Guo, Nie, et al., 2020; Guo, Sosa, et al., 2021; Di Persio, Tekath, et al., 2021; Shami et al., 2020; Sohni et al., 2019). Based on transcriptomic profiling of these subtypes, it has been found that the individual states do not directly correlate with type $A_{dark}$/$A_{pale}$/B spermatogonia as classified based on nuclear morphology. For instance, marker genes from all these subtypes show overlap with $A_{dark}$ spermatogonia at different proportions (Di Persio, Tekath, et al., 2021; Shami et al., 2020). Furthermore, laser capture micro-dissection of $A_{dark}$ and $A_{pale}$ spermatogonia reveals comparable transcriptomes (Jan et al., 2017).

State 0 undifferentiated spermatogonia in the adult human testis has been suggested to be the most naive stem cell population due to its resemblance to germ cells from the infant testis (Guo, Grow, Mlcochova, et al., 2018; Tan et al., 2020) and its ability to form spermatogonial colonies in germ cell transplantation assays (Shami et al., 2020; Tan et al., 2020). This is characterized by the expression of *PIWIL4*, *EGR4*, *C19ORF84* and *TSPAN33* (Di Persio and Neuhaus, 2023). Recently, two more subpopulations of State 0, termed State 0A and State 0B, have been identified (Sohni et al., 2019) and (Di Persio, Tekath, et al., 2021). These clusters fall on either side of

State 0 in the UMAP of scRNA-seq studies as cells commit to differentiation (Sohni et al., 2019).

**Spermatogenesis across development and ageing** Apart from characterizing the adult testis, researchers have also followed changes in healthy spermatogenesis across embryonic (Guo, Nie, et al., 2020; Garcia-Alonso et al., 2022), fetal (Guo, Sosa, et al., 2021), neonatal (Sohni et al., 2019), pre-pubertal development (Guo, Nie, et al., 2020), pubertal (Guo, Nie, et al., 2020) and ageing testis (Nie et al., 2022).

These studies have found that, at 14 weeks post fertilization, human male PGCs repress their pluripotency signatures such as *POU5F1* and *NANOG* and their early PGC markers such as *SOX17* and *TFAP2C* and turn on State 0 markers (*PIWIL4, MSL3, EGR4, TSPAN33*) to become State f0. State f0 cells persist through neonatal (Sohni et al., 2019) and infant testis (Guo, Sosa, et al., 2021). Following that, testes from 1- and 7-year-old boys mainly consist of State 0 and State 1 undifferentiated spermatogonia. However, the spermatogonia only form 3-4% of the testicular cells. From 11-year onwards, spermatogonial cells show both an increase in proliferation, forming ∼10-15% of the testicular cells, and in differentiation and commitment to meiosis. Finally, at 14-years, the testis resembles the adult stage and shows nearly full spermatogenesis (Guo, Nie, et al., 2020). During ageing, spermatogonia from older men only show modest transcriptional changes and a higher variation in colonizing potential compared to that of young adults (Nie et al., 2022).

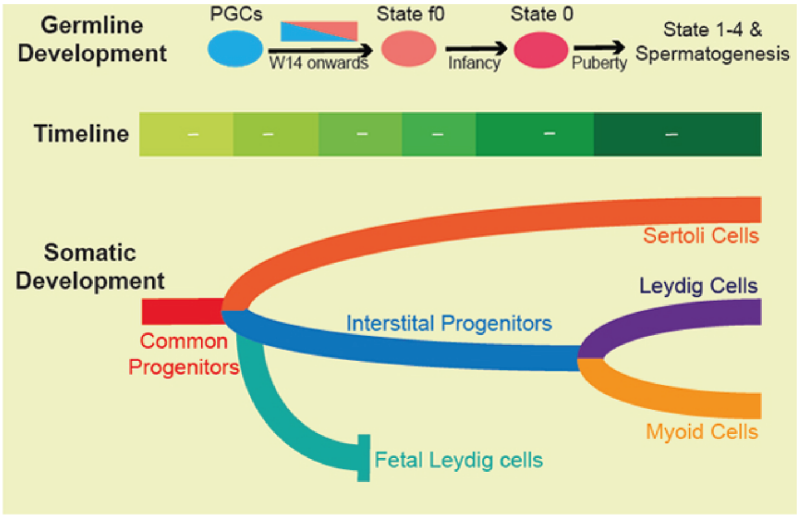



Figure 1.12: ***The human testis from the fetus to the adult (Guo, Sosa, et al., 2021).***

Within the somatic compartment, Garcia-Alonso et al., 2022 have identified a population of bipotent precursors in early gonads at 6-8 weeks post conception which become the first cells to express the testis-determining factor *SRY* in males. They further describe two testis-specific fetal macrophages (ftMs) including osteoclast-like *SIGLEC15*+ ftMs and microglia-like *TREM2*+ ftMs with possible functions in testis cord formation and minimizing inflammation. At 6-7 weeks post fertilization, a heterogeneous multipotent somatic cell progenitor exists that diverges into interstitial (including Leydig) and Sertoli lineages between 7-8 weeks. Around the same time at 7 weeks, fetal Leydig cells arise that persist throughout fetal development and become undetectable in the postnatal and infant stages (Guo, Sosa, et al., 2021). During

interstitial lineage development, the progenitors give rise to bipotent Leydig/myoid progenitor cells in the postnatal and prepubertal testis (Guo, Sosa, et al., 2021). At puberty, this bipotent population diverges to form separate lineages of Leydig and Peritubular Myoid cells (Guo, Nie, et al., 2020). During Sertoli cell development, the pre-pubertal testis consists of two immature states of Sertoli cells that differ in their metabolic and cell-cycle properties, and possibly represent alternative physiological states. The two states merge to form the mature Sertoli cells found from 11-year-old onwards (Guo, Nie, et al., 2020). Another study has identified Sertoli cells stages but propose them to be three consecutive stages of development (Zhao et al., 2020). In the ageing testis, the somatic compartment experiences a strong shift including an overall increase in the inflammatory response. Sertoli cells decrease in number and exhibit a dysregulation of the lipid metabolic pathway and energy production. Leydig cells further show decreased testosterone production and increased signatures of de-differentiation towards the common Leydig/myoid progenitor. The peritubular myoid cells exhibit hyperplasia and reduced contractility. These changes are more pronounced in aged men with higher body mass index. Altogether, the changes in the somatic compartment could impair germ cell development (Nie et al., 2022).

**Evolution of the testis transcriptome** Researchers have also availed single cell transcriptomic datasets to examine the rapid evolution of transcripts observed within testes (Brawand et al., 2011; Cardoso-Moreira et al., 2019; Wang, Leushkin, et al., 2020; Ramm et al., 2014). Multiple studies have found that the transcriptome consists of an increasing number of new genes during germ cell development, particularly in round and early spermatids (Murat et al., 2023; Shami et al., 2020; Lau et al., 2020). In the mouse testis, long terminal repeats (LTRs) of endogenous retroviruses (ERVs) have been shown to drive the expression of many novel transcripts that contain short open reading frames and produce detectable peptides (Davis et al., 2017). Additionally, during evolution, the X chromosome has accumulated genes that have specific expression in spermatogonia and Sertoli cells. X-linked genes required for post-meiotic function have also undergone repeated duplication to allow for active transcription and escape from MSCI (Murat et al., 2023).

The rate of gene evolution is also proposed to be regulated by a “transcriptional scanning” mechanism (Xia et al., 2020), whereby widespread transcription during spermatogenesis facilitates germline DNA repair (Hanawalt and Spivak, 2008; Werner et al., 2015) to reduce its mutation rates and, ultimately, the rates of gene evolution.

**Cross-species comparisons** Generating similar datasets from other organisms has allowed comparative studies across multiple species. State 0 spermatogonia are believed to be found only in the human testis. The most undifferentiated stem cell population in mouse testis is widely considered to be *Gfra1*+, which is equivalent to human State 1 spermatogonia (Nakagawa et al., 2021). However, Shami et al., 2020 and Carrieri et al., 2017 have identified very limited number of *Miwi2*+ spermatogonia in the mouse testis. The undifferentiated spermatogonia in humans (and macaque) are also present in higher numbers and undergo fewer mitotic divisions than in mouse. When considering species-specific gene expression, the spermatogonia show a greater divergence for transcription factors (TFs) than non-TF genes

Shami et al., 2020.

During meiosis, consistent with previous cytological data (Vries et al., 2012; Turner, Mahadevaiah, Fernandez-Capetillo, et al., 2005), scRNA-seq captures incomplete MSCI in human and macaque compared to mice (Shami et al., 2020). Human germ cells are further reported to proceed more quickly through the initial stages of meiosis compared to other eutherian mammals (Shami et al., 2020). Murat et al., 2023 extend this comparison to other mammalian orders to confirm that MSCI, contrary to previous observations (Daish et al., 2015), also occurs in monotremes and hence is common to mammalian sex-chromosome systems. Primate-specific transcripts, mainly consist of ERV-binding KRAB zinc-finger proteins (Imbeault et al., 2017), also show peak expression at meiosis (Shami et al., 2020).

In general, fertility-associated genes show conserved trajectories during spermatogenesis across species. Even ortholog genes show patterns of diverging trajectory across spermatogenesis. For example, *RUBCNL*, an autophagy gene (Cheng and Sun, 2019), has relatively reduced expression in human spermatids compared to that of Old World monkeys. Thus, it potentially plays distinct roles during spermatogenesis in the two lineages of species (Murat et al., 2023).

Testicular somatic cells also show differences across species. Peritubular myoid cells, which are present in multiple layers in primates and only in a single layer in mouse (Maekawa et al., 1996), show species-specific signalling (Shami et al., 2020). Leydig and myoid lineages are specified in the fetal testis in mice (Svingen and Koopman, 2013) but only during puberty in humans (Guo, Nie, et al., 2020). Lastly, putative germline-soma interactions involve different source and target cells across the different species (Shami et al., 2020; Guo, Grow, Mlcochova, et al., 2018).

**Altered spermatogenesis** Finally, sc/snRNA-seq approaches have supported a detailed examination of the tissue during altered spermatogenesis. Testosterone-suppressed transfemales were shown to have an absence or much lower proportion of germ cells from differentiating spermatogonia onwards, and a partial reversal of Sertoli cell developmental state (Guo, Nie, et al., 2020). Cryptozoospermia is a condition where exhibit a severe form of oligoasthenoteratozoospermia, i.e., defects in the morphology, number and motility of sperm, and fewer than 0.1 million sperm in the ejaculate (Tüttelmann et al., 2018). Sperm can be found after extended centrifugation and when viewed under the microscope (Baldi et al., n.d.). Testes from such patients have been found to show an increased proportion of *PIWIL4*+ undifferentiated spermatogonia and reduced number of $A_{dark}$ spermatogonia. They also exhibit prolonged expression of the transcription factors *EGR4*, which is involved in mouse germ cell differentiation (Hogarth et al., 2010) and *HOXC9*, which is involved in neuronal differentiation (Wang, Choi, et al., 2013; Di Persio, Tekath, et al., 2021). Non-obstructive azoospermia (NOA) is a more severe form of infertility, where no sperm is found even after extended centrifugation of the semen (Zhu et al., 2016). These patients were reported to have downregulated WNT signalling in the spermatogonia compartment. In general, NOA patients display a higher number of oxidative stress-associated genes in their somatic compartment (Wang, Liu, et al., 2018; Zhao et al., 2020). In idiopathic NOA, Sertoli cells are found in an immature state accompanied by increased WNT signalling (Zhao et al., 2020). In NOA that occurs with a known genetic disorder

in sex chromosomes, mature Sertoli cells are found, however, they show a chronic immune response (Zhao et al., 2020).

## 1.4 Gene regulation during human spermatogenesis

### 1.4.1 Chromatin accessibility

Exploration of the gene regulatory mechanisms that control the carefully orchestrated sequence of germ cell development is an upcoming area of research. One of the most fundamental mechanisms of gene regulation is through changes in chromatin accessibility. Linear DNA is tightly wound around histone proteins that form the nucleosomal structures. The DNA needs to become accessible for transcription factors to bind and induce the expression of their target genes. Chromatin accessibility can be measured through ATAC-seq (Assay for Transposase Accessible Chromatin with high-throughput sequencing). ATAC-seq involves the digestion of DNA by the Tn5 Transposase enzyme, which binds to and cuts only nucleosome-free open regions of the genome, and subsequent sequencing of these fragments (Buenrostro et al., 2015). ATAC-seq can be used to examine motif enrichment in accessible regions of the genome (Yan et al., 2020), infer putative enhancer associations through co-accessibility of peaks (Fullard et al., 2018) and even deduce TF binding strength through footprinting (Bentsen et al., 2020).

#### 1.4.1.1 snATAC-seq/multiome snATAC-seq + snRNA-seq technology

Modifications of the droplet-based sc/snRNA-seq technology have enabled the measurement of chromatin accessibility at the level of single cells. It is even possible to capture joint measurements of genome accessibility landscape and the accompanying gene expression patterns from the same single cells.

The technology involves the transposition of DNA within intact and permeabilized nucleus in solution prior to single cell partitioning. Gel beads within the individual droplets carry barcodes for the transposed DNA fragments, and in case of joint profiling of ATAC and RNA, additionally carry tags for the latter (*https://www.10xgenomics.com/* n.d.). Thus, joint measurement of ATAC-seq and RNA-seq provides a one to one correspondence between the chromatin accessibility landscape and gene expression patterns within a cell. Since its availability in the last few years, snATAC-seq/snATAC-seq + snRNA-seq has also been used to investigate the developmental process of spermatogenesis.

#### 1.4.1.2 Key findings from snATAC-seq of human spermatogenesis

**Stage-specific motif enrichment** While bulk ATAC-seq studies have not found any differences in chromatin accessibility profiles between undifferentiated and differentiating spermatogonia (Guo, Grow, Mlcochova, et al., 2018), snATAC-seq examination of the adult human testis has revealed stage-specific accessibility. The number of accessible peaks increases from spermatogonia to meiosis and decreases in spermatids. Motifs enriched in accessible peaks include DMRT1 in differentiating spermatogonia, SL1 and GLIS3 in pachytene/diplotene clusters, NFY from meiosis

onwards and CTCF and CTCFL/BORIS in all germ cells except undifferentiated spermatogonia (Wu et al., 2022).

## 1.4.2 Hierarchical organization of the 3D genome

The arrangement of the chromatin in 3D space is another mechanism of gene regulation. The human genome comprises 2m of DNA intricately folded in 3D space to fit into a 10 $\mu$m wide nucleus. This organization provides an opportunity for regions that are far apart in the linear genome to interact with each other in close proximity. The 3D organization of the genome has been shown to regulate or be associated with multiple biological processes including embryonic development (Hug and Vaquerizas, 2018), exit from pluripotency (Pękowska et al., 2018) and dosage compensation (Engreitz et al., 2013). In many instances, the disruption of this arrangement has also been associated with developmental disorders and disease (Lupiáñez et al., 2015).

The 3D organization of the genome is multi-layered and can be studied at various scales, from chromosomal territories to chromatin loops (Figure 1.13).

**Chromosomal territories** Individual chromosomes in interphase nucleus have been shown to preferentially fold into separate regions of the nucleus defined as chromosomal territories. This results in a higher frequency of interactions between different regions that reside on the same chromosome, than those that reside on different chromosomes (Cremer and Cremer, 2010).

**Compartments** Inter- and long-range intra-chromosomal interactions result in plaid patterns representing "compartmentalization" of the genome. This pattern results from the segregation of the genome into "A" and "B" compartments and the preferential inter-compartment interactions through possible phase separation mechanisms (Ling et al., 2022; Fujishiro and Sasai, 2022). Generally, the "A" compartment is found to contain actively transcribed genes and associated histone marks whereas the "B" compartment is found to contain inactive genes and modifications indicative of a repressed genomic state (Lieberman-Aiden et al., 2009; Kalhor et al., 2012; Sexton et al., 2012).

**Self-interacting domains** TADs or topologically associating domains are self-interacting domains of 100 kb-1 Mb that arise from locally enriched interactions of the chromatin (Dixon et al., 2012; Nora et al., 2012). TADs show an enrichment for active transcription and binding of CTCF and cohesin at their boundaries (Dixon et al., 2012).

**Loops** Loops are interactions that occur between specific non-neighbouring regions of the genome, like promoter-enhancer contacts (F. Jin et al., 2013; Schleif, 1992). Loops can also be formed between co-regulated genes (Schoenfelder, Sexton, et al., 2010), Polycomb-repressed genes (Schoenfelder, Sugar, et al., 2015), etc.

### 1.4.2.1 *in situ* Hi-C

Hi-C is a chromosome conformation capture (3C) technique which captures the interaction frequency between all pairs of DNA loci inside the nucleus (Figure 1.14).

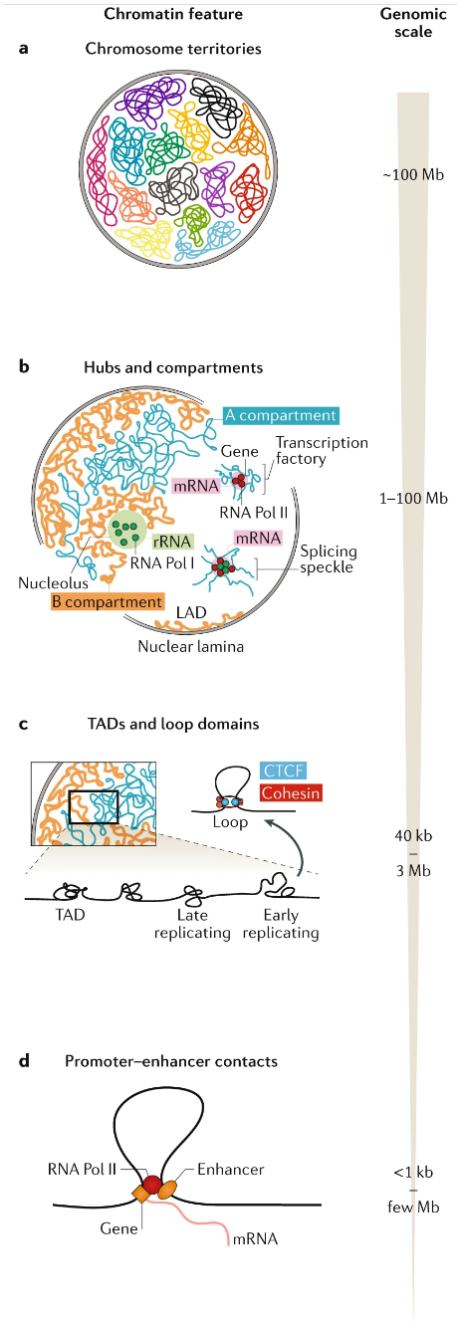



Figure 1.13: ***Hierarchical organization of the 3D genome (Kempfer and Pombo, 2020).***

The main principle behind 3C technology is the ligation of DNA fragments that are in close spatial proximity. In *in situ* Hi-C, the DNA within the nucleus is cross-linked using formaldehyde to stabilize DNA-DNA interactions that result from the folding of the chromatin in 3D space. Restriction enzymes are then used to digest the DNA and generate fragments with dangling ends. The dangling ends are filled with biotinylated nucleotides. In the subsequent proximity ligation step, ends of the DNA fragments are ligated to each other using blunt-end ligation (Lieberman-Aiden et al., 2009). Ligation is performed under diluted conditions and within the still intact nucleus. Hence, ligations between cross-linked DNA fragments are favoured over random ligations. The nucleus is then lysed and the biotinylated chimeric fragments are then pulled down and sequenced. The two pieces of interacting DNA in the chimeric fragment are in-

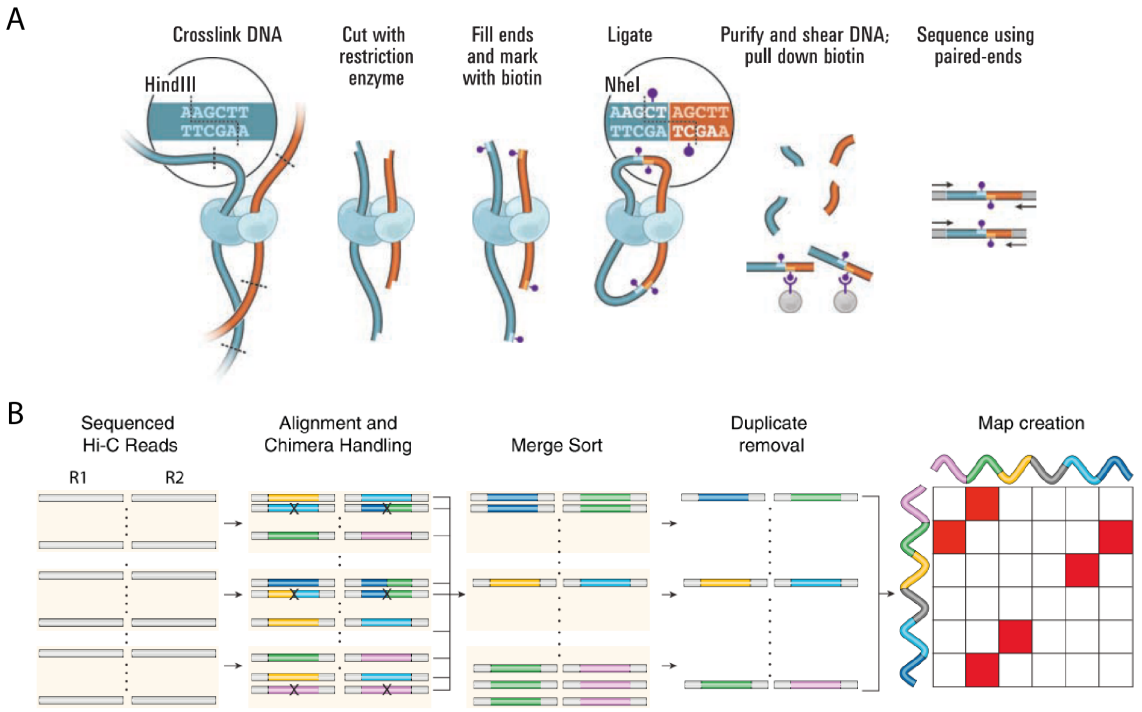



Figure 1.14: ***in situ Hi-C protocol and computational processing of the output. (Lieberman-Aiden et al., 2009; Durand et al., 2016).***

dependently aligned to the genome. Unpaired fragments, self-ligations and PCR duplicates are filtered out. The remaining fragments are used to produce an interaction frequency matrix that is often depicted as a heatmap.

The Hi-C heatmap which contains the normalized number of contacts between any two loci in the genome captures the multilayered organization of the 3D genome at different resolutions (Lieberman-Aiden et al., 2009; Belton et al., 2012) (Figure 1.15). Chromosomal territories appear as blocks of interactions per chromosome at hundreds of megabases along the diagonal. Compartments appear as plaid-patterns at 0.1-2 Mb that extends across chromosomes (Lieberman-Aiden et al., 2009), TADs occur as smaller blocks of high interaction frequency at 100 kb-1 Mb and loops appear as point interactions between different regions of the genome (Rao, Huntley, et al., 2014).

**Micro-C** Micro-C is an adaptation of the conventional Hi-C protocol, where the DNA is digested using micrococcal nuclease (MNase) instead of a restriction enzyme. This produces contact frequency heatmaps at nucleosomal resolution (Krietenstein et al., 2020; Hsieh, Cattoglio, et al., 2020). Additionally, Micro-C uses a second cross-linker such as disuccinimidyl glutarate (DSG) or ethylene glycol bis(succinimidylsuccinate) (EGS) which diminishes the number of random ligation events and improves the signal-to-noise ratio (Akgol Oksuz et al., 2021).

### 1.4.2.2 Key findings from 3D genome analysis of spermatogenesis

Since spermatogenesis exhibits extensive rearrangement of chromosomes, the characterization of its 3D chromatin architecture has been a highly active research area.

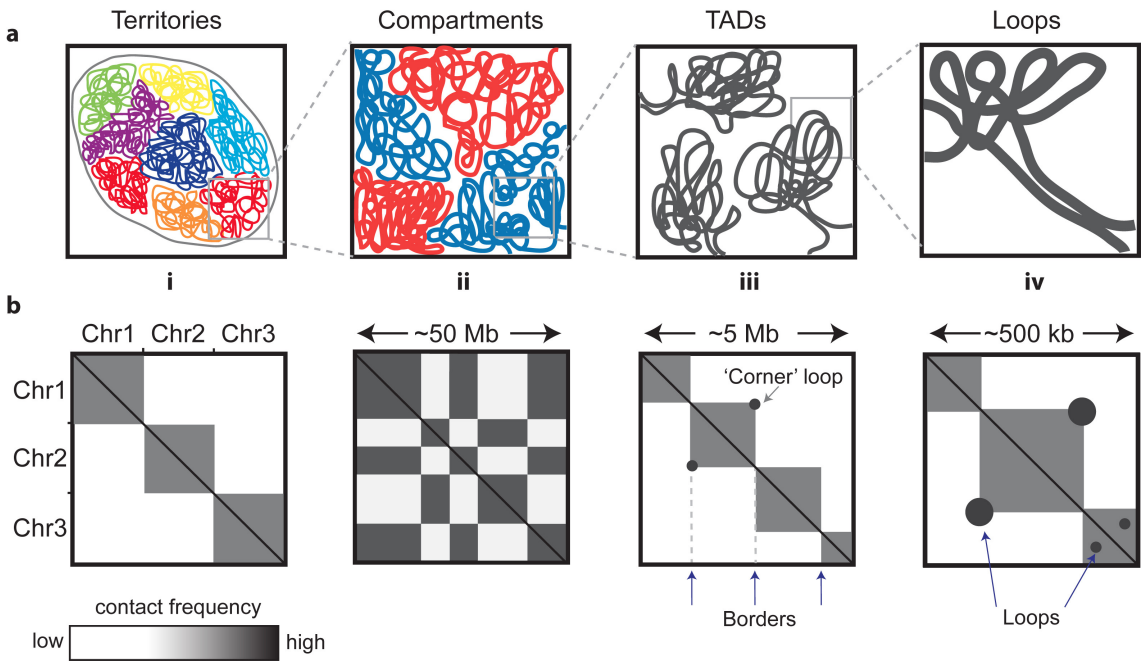



Figure 1.15: ***Representation of the hierarchical organization of the genome in the Hi-C heatmap (ssikorska2020defining).***

2019-2020 saw a boom of studies investigating the changes in 3D genome architecture during spermatogenesis and its interplay with transcription, cohesin/CTCF binding and the chromatin accessibility (Alavattam et al., 2019; Wang, Wang, et al., 2019; Luo et al., 2020; Patel et al., 2019; Vara et al., 2019). Due to the small size of the human testis biopsies, and the lack of an efficient system to enrich for the individual germ cell stages at high resolution and generate 3D genome maps from limited cell numbers, these studies are primarily in mouse and rhesus monkey. Key findings from these papers are summarized below.

**Compartments** During spermatogenesis in mouse and monkey, the spermatogonia exhibit strong compartmentalization which reduces to a minimum in meiotic pachytene spermatocytes undergoing recombination. The compartments re-appear in mature sperm, either to the same strength as present in spermatogonia or to a reduced level (Patel et al., 2019; Wang, Wang, et al., 2019; Vara et al., 2019). During differentiation from spermatogonia to pachytene spermatocytes, regions involved in double strand breaks (DSBs), cilia formation and piRNA clusters are reported to switch from the "B" to the "A" compartment (Luo et al., 2020). Mouse pachytene spermatocytes still present attenuated compartmentalization, and this organization includes interactions between synapsed homologous chromosomes (Patel et al., 2019). The attenuated compartments correlate with transcription (Alavattam et al., 2019) and show an enrichment of DSB hotspots (Patel et al., 2019). In mouse, inter-chromosomal interactions between the acrocentric ends (closer to the centromere) and between the non-centromeric ends of the different chromosomes are also visible in pachytene spermatocytes. This arrangement is consistent with microscopic observations of the meiotic prophase where the telomeres appear tethered to the nuclear membrane (Alavattam et al., 2019; Scherthan, 2001; Scherthan et al., 2014).

While the autosomes exhibit attenuated compartmentalization in mouse and mon-

key pachytene spermatocytes, the X chromosome, which undergoes MSCI at this stage, exhibits a complete lack of compartmentalization (Alavattam et al., 2019). Interestingly, reports of compartmentalization around MSCI-escaping genes are inconsistent. Vara et al., 2019 observe compartmentalization in mouse leptotene/zygotene cells around MSCI-escaping amplicons, while other groups report a complete absence of compartmentalization around these regions in monkey pachytene spermatocytes (Wang, Wang, et al., 2019) and in mouse round spermatids (Alavattam et al., 2019).

In rhesus monkey, the reduction in conventional compartments in pachytene spermatocytes is accompanied by a simultaneous emergence of a more refined-A/B segregation that correlates with transcription (Wang, Wang, et al., 2019). Refined-A/B is lost in mice that are mutant for *Sycp2* (part of the synaptonemal complex) or *Top6lb* (topoisomerase that interacts with SPO11 during DSBs to reduce torsional stress). However, the refined-A/B arrangement, is surprisingly not lost upon inhibition of transcription. The observed loss of refined-A/B in the *Sycp2*- and *Top6lb*-mutant pachytene spermatocytes is accompanied by the restoration of conventional compartments. Since mutants for different complexes show similar effects, the authors hypothesize that the meiotic stages are more important than the particular complexes present at each of the stages in determining the chromatin structure.

**Self-interacting domains** Mouse and monkey spermatogonia exhibit typical TAD structures, which become larger ($\sim$1.1 Mb in mouse) and weaker in pachytene spermatocytes. TADs return to their stronger and smaller arrangement in round spermatids (Vara et al., 2019). While most studies report a resurgence of conventional TADs ($\sim$0.56 Mb) in mouse and monkey mature sperm (Patel et al., 2019; Wang, Wang, et al., 2019; Alavattam et al., 2019; Luo et al., 2020), Vara et al., 2019 find an absence of the same in mouse. The authors of Vara et al., 2019 attribute the difference in their observation to the purer populations of spermatogenic cells obtained in their study through fluorescence-associated cell sorting (FACS) compared to the commonly used STA-PUT velocity sedimentation technique employed in the other studies. Surprisingly, the loss of TADs in pachytene spermatocytes arises even though these cells exhibit CTCF and RAD21 binding to the same levels and at the same boundaries as seen in mouse spermatogonia (Patel et al., 2019; Vara et al., 2019).

As with compartmentalization, the X chromosome does not exhibit any TADs during MSCI in mouse and monkey pachytene spermatocytes (Alavattam et al., 2019). However, the X chromosome show specific off-diagonal perpendicular signal that could indicate its folding within the XY body (Wang, Wang, et al., 2019). In this regard, MSCI differs from female random X chromosome inactivation (XCI). XCI is a dosage compensation mechanism where one of the two copies of the X chromosome in therian female mammals is silenced and packaged into a unique 3D structure called the Barr body (Barr and Bertram, 1949). While the silenced chromosome in XCI still shows a bipartite domain at *Dxz4* (Darrow et al., 2016), MSCI exhibits no such domain structure. Similar to conventional compartmentalization, the lost TADs in wild type pachytene spermatocytes are also restored in mouse pachytene spermatocytes mutant for *Sycp2* or *Top6lb* (Wang, Wang, et al., 2019).

In studies that do observe a restoration of TADs in mouse mature sperm, the domain

boundaries are not enriched for histones retained after the histone-to-protamine exchange (Luo et al., 2020). However, the TAD boundaries in mature sperm already show a deposition of H3K4me3, H3K27ac and H3K27me3 at the earlier stages of pachytene spermatocytes and round spermatids (Alavattam et al., 2019). Thus, the authors hypothesize that epigenetic mechanisms that persist through the meiotic divisions could determine the 3D chromatin organization in the germline.

**Loops** Reports of changes in looping interactions during spermatogenesis are conflicting. While some studies cite an absence of loops in mouse and monkey pachytene spermatocytes (Luo et al., 2020; Wang, Wang, et al., 2019), some others find pronounced looping interactions that overlap active transcription, H3K4me3 and H3K27ac marks in mouse (Alavattam et al., 2019). These loops extend in size from 0.8-1 Mb in zygotene to 1.5-2 Mb in pachytene spermatocytes when synapsis stabilization occurs (Patel et al., 2019). The loops regained in mouse mature sperm are further enriched in genes expressed in the embryo (Luo et al., 2020).

**3D genome in human mature sperm** Surprisingly, the human mature sperm was reported to completely lack TADs, even while showing strong A/B compartmentalization. The loss of TADs was attributed to the lack of CTCF in human sperm (Chen, Ke, et al., 2019).

**Confounding effects in studies of sperm DNA** In recent times, multiple papers (Galan et al., 2021; Chen, Scott-Boyer, et al., 2022; Yin et al., 2022) have started questioning the purity of the sperm DNA analysed in published studies. Yin et al., 2022 propose that the observations of TADs in mouse mature sperm have been confounded by the presence of contaminating cell-free DNA in the sperm obtained from the epididymis. This is compounded by the incomplete lysis of the sperm head leading to a lack of accessibility to pure sperm DNA. While repeating published experiments of Hi-C in mouse mature sperm after the removal of cell-free DNA with DNAse I and efficient lysis of the sperm head with a strong reducing agent like DTT prior to detergent treatment, the authors find that the mouse sperm no longer show compartments (except extremely weak signals) or TADs. Repeating ChIP-seq with the inclusion of these modifications to the protocol also results in a lack of CTCF binding to sperm genome, though the authors are unsure whether this is truly biological or if the additional treatment has adversely affected the binding epitope of the CTCF antibody.

The authors further hypothesize that Vara et al., 2019's observation of the lack of TADs in the murine mature sperm could be due to the usage of DNAse I in their testis digestion protocol, and due to breakage of sperm head from snap-freezing of FACS-sorted cells prior to Hi-C library generation. They also lean towards the observed lack of TADs in human sperm as seen in Chen, Ke, et al., 2019, even though the study did not include DTT treatment. The authors reason that the human sperm is more readily permeabilized than the mouse sperm due to a higher histone content in the human sperm genome and a lower cysteine content in the human protamines (Hisano et al., 2013). Hence, they believe that the human sperm does not necessarily need DTT for successful lysis. The authors also reason that the possible lower levels of cell-free DNA in the ejaculated sperm (as used in Chen, Ke, et al., 2019) could lead to lesser contamination and more accurate results.

Yin et al., 2022 question the fundamental methodologies employed in sperm DNA preparation and propose elegant explanations for the contextual observations made in the various studies so far. However, there is still no conclusive and consensual evidence for the 3D genome architecture in mature sperm.

## 1.5 Aims of the thesis

As seen from the studies mentioned earlier, transcriptomic characterization of the adult human testis has provided great insights into the process of spermatogenesis. However, little is known about the gene regulatory mechanisms that control this complex process. Thus, in this thesis, we are employing state-of-the art genomic techniques to characterize the dynamics of chromatin accessibility and the three-dimensional organization of the genome that drive adult human spermatogenesis.

In particular, the main aims of the thesis include:

1. Generation and characterization of joint single nuclear chromatin accessibility (snATAC-seq) and transcriptomic (snRNA-seq) multiome datasets of the adult human testis
2. Investigation of the dynamics of spermatogonia homeostasis in the adult human testis
3. A systematic comparison of chromosome X- or Y-bearing haploid spermatids in the adult human testis
4. Generation and preliminary characterization of high resolution maps of the 3D chromatin organization across key developmental stages during adult human spermatogenesis

# Chapter 2

# Methods

## 2.1 Selection of human testicular biopsies

We obtained testicular biopsies of patients from the Centrum für Reproduktionsmedizin und Andrologie (CeRA) in Münster, Germany. Normal spermatogenesis samples were obtained from patients who had been diagnosed with obstructive azoospermia due to a previous vasectomy, a congenital malformation or physical obstruction of the vas deferens, but had normal testicular volume, normal FSH levels and qualitatively and quantitatively normal spermatogenesis. A complete list of the measured clinical parameters can be found in the Appendix. When undergoing surgery for microdissection testicular sperm extraction (mTESE) or histological evaluation at the Department of Clinical and Surgical Andrology (University Hospital in Münster, Germany), one additional testicular sample was excised for this study after written informed consent. Ethical approval was obtained for this study (Ethics Committee of the Medical Association of Westphalia-Lippe and University of Münster 2010-578-f-S).

## 2.2 Preparation of single cell suspensions from human adult testis biopsy

A two-step enzymatic digestion was performed as detailed in Kossack et al., 2013; Laurentino et al., 2019; Neuhaus et al., 2017. Briefly, the tissue sample was placed into 6 cm culture dish with MEM-alpha media (Life Technologies GmbH, Gibco, Darmstadt, Germany) and dissected into fragments of 1mm or less using a blade. It was then transferred into a 15 ml tube and centrifuged for 5 minutes at 1500 rpm and 10℃. The cell pellet was resuspended in MEM-alpha media containing 1mg/ml filter sterilized collagenase IA (C9891, Sigma) and incubated at 37℃ until tubules separated. The tube was inverted every 5 minutes. The reaction was stopped by adding 5 ml MEM-alpha media with 10% FBS and 1% Pen/Strep. It was centrifuged at 10 ℃ at 1500 rpm for 4 minutes, the supernatant discarded and incubated with 2 ml filter sterilized Hank's balanced salt solution (Life Technologies GmbH, Gibco) containing 2 mg Trypsin (27250-018, Gibco) and 11 mg DNase I (DN25, Sigma) for 7-10 minutes at 37℃. The tube was inverted every 3 minutes. This was followed by strong pipetting to obtain a single-cell suspension. The reaction was stopped as outlined above

and cells were washed twice with MEM-alpha. Then, erythrocytes were removed by incubating the cells with haemolysis buffer (0.83% NH4Cl solution) for 3 minutes at room temperature. Enzymatic digestion was stopped as outlined above. The cell suspension was filtered through a 70 $\mu$m strainer to remove cell clumps and debris. It was pelleted, washed and the cells were resuspended in 2 ml medium prior to cell counting using the trypan blue exclusion method.

## 2.3 Generation of paired snATAC-seq and snRNA-seq libraries

200,000-400,000 cells were centrifuged at 500g for 10 minutes at 4℃. The supernatant was discarded, and the pellet resuspended in 50 $\mu$l Wash Buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1X Protease Inhibitor Cocktail). It was centrifuged at 500g for 10 minutes at 4℃ and supernatant was discarded. The pellet was incubated with 45 $\mu$l Lysis Buffer (wash buffer with 0.01% NP40, 0.01% Digitonin, 1mM DTT and 1 U/$\mu$l RNAse inhibitor (3335399001, Sigma)) for 25-30 seconds on ice. The reaction was stopped with 50 $\mu$l Neutralising Buffer (lysis buffer with 2mM EDTA). It was again centrifuged at 500g for 10 minutes at 4℃, the supernatant discarded, and the pellet washed with 45 $\mu$l 1X Dilute Nuclei Buffer (20X Nuclei Buffer provided with 10x scATAC + GEX multiome kit). The pellet was resuspended in 10 $\mu$l 1X Dilute Nuclei Buffer solution and 2 $\mu$l sample was taken for counting with Trypan Blue.

9000-10,000 nuclei were loaded for the 10x scATAC + GEX multiome transposase reaction and the library prepared according to manufacturer's instructions in Chromium Next GEM Single Cell Multiome ATAC + Gene Expression Reagent Kits User Guide (Revision A). The libraries were sequenced in NovaSeq 6000 with paired end sequencing.

RNA-seq libraries were sequenced to a depth of 40,000 read pairs/nucleus and ATAC-seq libraries to a depth of 50,000 read pairs/nucleus according to 10x sequencing guidelines.

## 2.4 snATAC-seq and snRNA-seq analysis

### 2.4.1 Read mapping and quantification

The fastq files of individual samples were processed with cellranger-arc v2.0.0 and mapped to refdata-cellranger-arc-GRCh38-2020-A-2.0.0.tar.gz. For each sample, cellranger-arc count was used with default parameters to generate a filtered peaks x cells counts matrix and a filtered genes x cells counts matrix. The filtered matrices of each sample were loaded into R and further processed with Seurat v4.1.1 (Hao et al., 2021) and Signac v1.7.0 (Stuart et al., 2021).

### 2.4.2 RNA library QC

Ambient RNA contamination was estimated and removed using soupX v1.6.1 (Young and Behjati, 2020). Next, doublets were removed using scDblFinder v1.10.0 (Germain et al., 2021), using a doublet rate of 0.077 or 0.069, corresponding to the multiplet rate for 10,000 and 9,000 recovered cells respectively in the 10x multiome scATAC + GEX user guide (Revision A). Cells with > 20% mitochondrial reads, fewer than 500 UMIs or fewer than 250 detected genes were removed. These cut-offs were determined using the $isOutlier()$ function from scater v1.24.0 (McCarthy et al., 2017) or by visually examining the distribution with the aim of retaining as many high quality cells as possible.

### 2.4.3 ATAC library QC

Doublets were removed with using scDblFinder v1.10.0 (Germain et al., 2021), using a doublet rate of 0.077 and 0.069 for 10,000 and 9,000 recovered cells respectively, and aggregateFeatures = TRUE. Cells with fewer than 150 detected peaks or fewer than 250 fragments were removed. These cut-offs were chosen as described above. Cells with fraction of reads in called peak regions (FRiP) < 0.20 or Transcription Start Site (TSS) enrichment score < 2 were removed.

### 2.4.4 Peak calling

Peaks were called on the remaining high quality cells grouped by Cellranger-identified cluster, using macs2 (Zhang, Liu, et al., 2008) via the Seurat function $CallPeaks()$. Peaks on non-standard chromosomes and those in genomic blacklist regions were removed. A new assay was created by quantifying counts in each of the remaining peaks with function $FeatureMatrix()$ from Seurat taken for all downstream processing.

### 2.4.5 Normalization, sample integration, dimensional reduction

The RNA and ATAC-seq assays of the biological replicates were separately integrated using Seurat.

#### 2.4.5.1 RNA-seq assay

Genes x cells matrices were normalized with scran v1.24.0 (Lun et al., 2016) as follows. The Seurat object was converted to a SingleCellExperiment (SCE) object and clustered using the $quickCluster()$ function. Size factors were calculated using $computeSumFactors()$. The $logNormCounts()$ function from scuttle v1.6.2 (McCarthy et al., 2017) was used to calculate normalized counts with log = FALSE. The normalized counts were then log-transformed with the $log1p()$ function and added to the data slot of the Seurat object.

The top 4000 variable features were identified with Seurat's $FindVariableFeatures()$ function with selection.method = "vst". The datasets were then integrated using the Seurat functions $SelectIntegrationFeatures()$, $FindIntegrationAnchors()$, and

$IntegrateData()$. Finally, the combined integrated matrix was then scaled and centred with function $ScaleData()$ from Seurat and Principal Component Analysis (PCA) was performed using $RunPCA()$.

#### 2.4.5.2 ATAC-seq assay

The union of peaks called in each individual snATAC-seq dataset was taken, and all datasets were re-quantified using this unified peak set. The bottom 5% of features were removed and Term Frequency - Inverse Document Frequency (TF-IDF) normalization and singular value decomposition (SVD) were carried out to produce an latent semantic indexing (LSI) matrix. Datasets were then merged and the LSI calculation repeated. The LSI embeddings were integrated using the functions $FindIntegrationAnchors()$ (using all peaks and reciprocal LSI) and $IntegrateEmbeddings()$.

### 2.4.6 Weighted nearest neighbour (WNN) identification, clustering, and cluster annotation

The Seurat function $FindMultiModalNeighbors()$ was used on the integrated RNA data (PCA dimensions 1:20) and ATAC data (LSI dimensions 2:20) with k.nn = 25 for the normal samples. Clustering was carried out on the resulting WNN network using the $FindClusters()$ function with res = 0.8 for normal samples. Clusters were annotated based on the expression of key marker genes (Di Persio, Tekath, et al., 2021; Di Persio and Neuhaus, 2023).

## 2.5 scRNA-seq analysis for scRNA vs snRNA comparison

### 2.5.1 Read mapping

The fastq files of individual samples were processed with cellranger v6.0.1 and mapped to refdata-gex-GRCh38-2020-A.tar.gz. For each sample, cellranger count was used with default parameters to generate a filtered genes x cells counts matrix.

### 2.5.2 Quality control filtering

The filtered matrices of each sample were loaded into R and further processed with Seurat v4.0.3. Doublets were removed using scDblFinder v1.10.0 with the following doublet rates: 0.008 for Guo, Grow, Mlcochova, et al., 2018, 0.046 for Di Persio, Tekath, et al., 2021, and 0.077 or 0.069 for 10,000 or 9,000 cells recovered in this study, according to the respective 10x user guides. Cells with > 24% mitochondrial reads, nFeature_RNA < 350 (< 1000 for this study) and > 9500 were removed. These cut-offs were based on function $isOutlier()$ from scater v1.20.1 or by visually examining the distribution with the aim of retaining as many high quality cells as possible. Sample N1 from Di Persio, Tekath, et al., 2021 was excluded from the analysis due to high ambient RNA contamination that resulted in higher than estimated number of droplets being called as cells by Cellranger.

Each dataset was normalized with scran v1.24.0 as described above for multiome data. The top 4000 variable features were identified with Seurat function $FindVariableFeatures()$ with selection.method = "vst".

### 2.5.3 Normalization, sample integration, dimensional reduction

Datasets were then integrated with Seurat using the Seurat functions $SelectIntegrationFeatures()$, $FindIntegrationAnchors()$, and $IntegrateData()$, with normalization.method = "Log-Normalize". The integrated assay was scaled and centred with Seurat function $ScaleData()$ with all genes as features and PCA was performed using $RunPCA()$.

### 2.5.4 Clustering and cluster annotation

Cells were clustered using Seurat function $FindClusters()$ on the integrated RNA with dims 1:30 and resolution = 0.5. Clusters were annotated based on key marker gene expression, as used in Di Persio, Tekath, et al., 2021.

## 2.6 Subclustering of spermatogonia clusters

### 2.6.1 Normalization, sample integration, dimensional reduction

After subsetting the spermatogonia clusters with the function $subset()$, the RNA-seq and ATAC-seq libraries of the individual replicates were re-integrated. For this, the subset Seurat object was first split into the individual replicates.

#### 2.6.1.1 RNA-seq assay

All steps were repeated as before except the normalization. Since scran produces normalized values per cell, it is not necessary to repeat normalization after subsetting cells. For each replicate, the top 4000 variable features were identified and the replicates integrated. The integrated assay was scaled and centred and PCA was computed based on the integrated assay.

#### 2.6.1.2 ATAC-seq assay

All steps were repeated as before. For each replicate, the bottom 5% of features were removed and TF-IDF and SVD were carried out. Datasets were then merged and the LSI calculation repeated. The LSI embeddings were integrated using the functions $FindIntegrationAnchors()$ (using all peaks and reciprocal LSI) and $IntegrateEmbeddings()$.

### 2.6.2 Weighted nearest neighbour (WNN) identification, clustering, and cluster annotation

The Seurat function $FindMultiModalNeighbors()$ was used on the integrated RNA data (PCA dimensions 1:16) and ATAC data (LSI dimensions 2:25) with k.nn = 30 for normal samples. Clustering was carried out on the resulting WNN network using the $FindClusters()$ function with res = 1.2 for normal samples. Clusters were annotated

based on the expression of key marker genes used in Di Persio, Tekath, et al., 2021; Di Persio and Neuhaus, 2023.

## 2.7 Differential gene expression

Identification of cluster-specific enriched genes for the whole testis dataset and the spermatogonia subclusters was carried out using the $FindAllMarkers()$ function from Seurat, using method “MAST” (Finak et al., 2015) with biological replicate as a latent variable. Only positively enriched markers with an adjusted p value < 0.05 were kept.

Identification of differentially expressed genes between snRNA-seq and scRNA-seq and between X- and Y-spermatids was carried out for each cluster independently using Seurat $FindMarkers()$ with method "MAST" with biological replicate as a latent variable. Genes with an adjusted p value < 0.05 and an absolute log2 foldchange in expression of > 0 were kept. For X- vs Y-spermatids, differentially expressed genes were calculated after removing genes located in PAR regions.

## 2.8 GO term enrichment

For GO term enrichment in cluster-specific differentially expressed genes between snRNA-seq and scRNA-seq samples, all genes with BH-adjusted p value < 0.05 and an absolute log2 foldchange in expression of > 0 were used as input for GO term enrichment analysis with clusterProfiler (v4.7.1.003) function $enrichGO()$ (Yu, Wang, Han, et al., 2012), using all genes included in the differential gene expression analysis as the background gene set. Up- and downregulated genes were analysed separately. The results were simplified to remove terms with high semantic similarity using method = “Wang” and cutoff = “0.7”. For visualisation, the most significantly enriched (based on BH-adjusted p value) term per cluster comparison was selected and enrichment of these terms across all clusters plotted.

For spermatogonia subcluster analysis, the top 100 most significantly enriched genes per cluster were used as input and processed as described above.

## 2.9 Differential accessibility

For spermatogonia subcluster analysis, identification of cluster-specific accessible peaks was carried out using Seurat $FindAllMarkers()$ with method “LR”. The total number of reads in peaks was used as a latent variable and only peaks with signal in at least 5% of cells in the cluster were assessed. Peaks with an adjusted p value < 0.05 were kept.

Identification of differentially accessible peaks between X- and Y-spermatids samples was carried out after removing peaks overlapping pseudoautosomal regions for each cluster independently as described above. Peaks were annotated with ChIPseeker v1.32.1 (Yu, Wang, and He, 2015) function $annotatePeak()$.

## 2.10 Motif activity

Motif activity was calculated using the Arch v1.0.2 R (Granja et al., 2021) implementation of ChromVar (Schep et al., 2017). Cellranger output filtered matrices were pre-processed using ArchR, keeping cells with a minimum of 80 fragments and a minimum TSS enrichment value of 1.45. Cells were then further filtered to keep only those that had passed the filtering and quality control procedures described above. Consensus motifs from Vierstra et al., 2020 obtained from https://resources.altius.org/~jvierstra/projects/motif-clustering-v2.0beta/were used for motif activity calculations using ArchR with default parameters. The resulting ChromVar Z-scores were then added back to the integrated Seurat object for downstream analysis.

## 2.11 Differential motif activity

For spermatogonia subcluster analysis, identification of cluster-specific motif activity peaks was carried out by computing the average difference in chromVAR z-score using Seurat $FindAllMarkers()$ with mean function "rowMeans". Motifs with an adjusted p value < 0.05 were kept.

Identification of differential motif activity between X- and Y-spermatids samples was carried out Seurat $FindMarkers()$ for each cluster independently as described above. Motifs with an adjusted p value < 0.05 and absolute average difference > 0 was kept.

## 2.12 Correlation analysis

For snRNA-seq vs scRNA-seq, average gene expression was calculated for each cluster with Seurat function $AverageExpression()$. The correlation coefficient for average gene expression was calculated with corrplot v0.92 (Wei and Simko, 2021) function $cor()$ and method = "pearson". The p values were calculated with the function $cor.mtest()$.

For gene expression analysis of the spermatogonia subclusters, average gene expression was calculated for each cluster with Seurat function $AverageExpression()$ for the top 100 clusters markers with the highest log2 foldchange and adjusted p value < 0.05. The correlation coefficient for average gene expression was calculated with corrplot v0.92 function $cor()$ and method = "spearman". The p values were calculated with the function $cor.mtest()$ with method = "spearman" and exact = FALSE.

For motif activity analysis of the spermatogonia subclusters, average motif activity was calculated for each cluster with Seurat function $AverageExpression()$ for the top 100 cluster-specific motifs with the highest average difference and adjusted p value < 0.05. The correlation coefficient for average motif activity was calculated with corrplot v0.92 function $cor()$ and method = "spearman". The p values were calculated with the function $cor.mtest()$ with method = "spearman" and exact = FALSE.

## 2.13 TF footprinting

In order to carry out TF footprinting for each cluster, the subset-bam tool from 10x (v1.1.0) was used to produce bam files containing reads for cells assigned to each cluster. Individual replicates were merged using samtools v1.16.1 (Danecek et al., 2021; Bonfield et al., 2021) to produce one bam file per cluster. These bam files were used as input to TOBIAS v0.13.3 (Bentsen et al., 2020) for footprinting of motifs from Vierstra et al., 2020 obtained from https://resources.altius.org/∼jvierstra/projects/motif-clustering-v2.0beta/within peaks identified in any cluster from the Seurat object. Regions falling into the blacklist regions included in the Signac package were excluded.

TOBIAS PlotAggregate was used to aggregate TF binding scores in 120bp windows across all detected motifs within peaks. For each motif and cluster, the average signal in the central 20bp around the motif and average signal in the flanking regions were calculated. Motifs with a notable footprint have less signal in the central region than in the flanking regions (a positive footprint depth).

## 2.14 Cell cycle analysis

Cell cycle analysis was performed using Tricycle v1.4.0 (Zheng et al., 2022). The integrated RNA data was converted to a SingleCellExperiment (SCE) object and projected into a cell cycle embedding obtained from a pre-learned reference projection matrix using the Tricycle function $project\_cycle\_space()$ with the type of species set to "human" and type of gene names set to "SYMBOL". Then, the cell cycle score was estimated with the function $estimate\_cycle\_position()$ and added back to the metadata of the Seurat object.

To construct a UMAP after regressing out the cell cycle score, first, the integrated RNA data was re-scaled and centred with the Seurat function $ScaleData()$ with vars.to.regress set to the Tricycle score. Then, the PCA was re-computed using $RunPCA()$. The Seurat function $FindMultiModalNeighbors()$ was used on the new PCA (PCA dimensions 1:16) along with the integrated ATAC data (LSI dimensions 2:25) with k.nn = 30. The UMAP was then recomputed with the Seurat function $RunUMAP()$ with n.neighbors = 30.

## 2.15 Slingshot trajectory inference

Trajectory inference was carried out using Slingshot v2.4.0 22, using the first 16 dimensions of the PCA reduction of the RNA data after ambient RNA removal. The cluster corresponding to Spermatogonia State 0 was specified as a trajectory start point. The distance metric was set to the default value of "slingshot" or "mnn" and omega = TRUE.

## 2.16 MultiVelo analysis

For velocity analysis, firstly, the RNA data from the integrated Seurat object was exported to AnnData with Seurat functions $SaveH5Seurat()$ and then $Convert()$.

Independently, the spliced and unspliced mRNA count matrices for each biological replicate were constructed using velocyto v0.17.17 (La Manno et al., 2018) run. The resulting .loom files were concatenated in the same order of replicates as in the RNA AnnData object. The concatenated .loom file was merged with RNA AnnData using the scVelo v0.2.4 (Bergen, Lange, et al., 2020) function $scv.utils.merge()$. For the ATAC data, the Gene Activity matrix was calculated using the Seurat function $GeneActivity()$ and this was exported to AnnData as described above. Additionally, the neighbourhood information was exported as text files from the "weighted.nn" slot of the integrated Seurat object.

For MultiVelo (Li, Virgilio, et al., 2023) analysis, the RNA Anndata containing the spliced and unspliced counts, and the ATAC AnnData were loaded with the function $scv.read()$.

From the RNA AnnData, the top 1000 variable genes were calculated using the function $scv.pp.filter\_and\_normalize()$ with flavor set to "seurat" and the specific retention of the following genes of interest: *DDX4, MAGEA4, PIWIL4,EGR4, FGFR3, SERPINE2, NANOS2, RRBP1, USP11, KIT, DMRT1, MKI67, DNMT1, REC8, STRA8, SOHLH2*. For the ATAC AnnData, the aggregated peaks were normalized with the MultiVelo v0.1.2 function $mv.tfidf\_norm()$. Then, genes that are present in both the RNA and ATAC AnnData were selected as "shared genes". The RNA AnnData was re-loaded and both the RNA and ATAC AnnData were subset to contain only the shared genes.

The re-loaded and subset RNA AnnData was preprocessed again using $scv.pp.filter\_and\_normalize$ as before, followed by $scv.pp.moments()$ with 30 PCA dimensions and 50 neighbours. The subset ATAC AnnData was smoothened with the neighbourhood information exported earlier using the function $mv.knn\_smooth\_chrom()$. Clusters with label "Unclear" were removed and the MultiVelo dynamics recovered using $mv.recover\_dynamics\_chrom()$ with the embedding set to the WNN UMAP.

The velocity graph was computed based on the recovered dynamics using neighbourhood connectivities with $mv.velocity\_graph()$. The top genes for the phase portraits were chosen based on the values for fit_likelihood and fit_likelihood_c for spliced vs unspliced and unspliced vs chromatin accessibility respectively. Latent time was computed using $mv.latent\_time()$ with the root_key set to the index of a cell belonging to the cluster 1: State 0.

## 2.17 Cell-cell communication analysis

### 2.17.1 Predicting Ligand-Receptor Interactions

To determine possible ligand-receptor interactions (LRI) between adult human testicular cells, a permutation-based approach similar to the one introduced in CellphoneDB (Efremova et al., 2020) was used. The normalized count matrix (“data”) was used as

a basis for 10,000 celltype label permutations where cell type labels were randomly re-assigned. Ligand/Receptor (LR) interaction pairs were taken from the consensus database of LIANA (Dimitrov et al., 2022), excluding bidirectional annotations and genes expressed in less than 10% of cells. For each Ligand-Receptor interaction first the mean expression of the ligand (Lmean) and receptor (Rmean) were calculated for each cell type in each condition and then an interaction score (LRIScore) was calculated for all possible cell type interactions per condition separately as the product of Lmean and Rmean. This approach differs from the methodology employed by CellPhoneDB in that our approach uses the product instead of the mean value of Ligand and Receptor mean expression. This is intended to implicitly penalize interactions where one or both partners are lowly expressed. LRI-Score calculation was repeated for each permutation of the cell type labels, and it was assessed whether the newly calculated score was higher or lower than the original score. After performing all permutations in such a manner, a p-value was calculated as the number of values equal or higher to the non-shuffled value, divided by the number of permutations. Values of p calculated in this test are approximations of p, but will be referred to as p-values nonetheless. Multiple testing correction was performed using the Benjamini & Hochberg (Benjamini and Hochberg, 1995) implementation in the statsmodels (Seabold and Perktold, 2010) function multiple tests on all p-values. These corrected p-values will be referenced as adjusted p-values. Additionally, using the created background distribution, the z-score of the original LRI Score was determined. In the same fashion as for cell-type labels, condition labels (State 0A/State 0B) were permuted an equal number of times as cell type labels and a difference score between the LRI scores of the two conditions were calculated for each cell-type interaction as:

$$LRIDiff_{score} = \frac{log2(Lmean_{State0A}/Lmean_{State0B}) + log2(Rmean_{State0A}/Rmean_{State0B})}{2}$$

p-values and adjusted p-values were calculated from these scores in a manner sensitive to the direction of change and used to assess differences in LRI scores between conditions.

## 2.18 Conjugation of fluorophore to antibodies

The antibodies were conjugated using Dylight Microscale Antibody Labelling Kits (ThermoScientific; Germany). 100 $\mu$l of antibody was taken at 1 mg/ml concentration in sterile PBS. If the antibody was present in a different buffer or concentration, it was purified or concentrated using a Vivaspin 500 column with 10,000 MWCO (VS0102, Sartorius). Then 8 $\mu$L of the Borate Buffer (0.67M) was added to the purified and concentrated antibody and the prepared protein added to 1 vial of DyLight Reagent. It was vortexed gently and incubated for 60 minutes at room temperature protected from light. After placing a spin column in a collection tube, the purification resin was mixed uniformly, 100 $\mu$l added to the spin column and centrifuged for 1 minute at 1000g to remove the storage solution. After discarding the collection tube, the spin column was placed into a new tube. Then, 108 $\mu$L of the labelling reaction was added to the column and allowed to mix with the resin by briefly vortexing. It was then centrifuged for 1 minute at 1000g to collect the purified protein. The absorbance was measured at 280 nm and Amax of the specific dye in Nanodrop 8000. These values were used to cal-

culate concentration and degree of labelling according to manufacturer's instructions. The labelled protein protected from light and stored at 4℃.

## 2.19 Whole mount staining of human seminiferous tubules

Dehydrated tubules from human testis biopsies were placed in a 24-well plate at room temperature and sequentially incubated with agitation in 50% Ethanol for 10 minutes, 30% Ethanol for 10 minutes, distilled water for 10 minutes, TBS for 10 minutes and 1M glycine for 1h. The tubules were incubated with blocking solution (2% BSA, 5% Normal Donkey Serum, 0.2%Triton X-100 in TBS) over night at 4℃. For triple staining with fluorophore-conjugated antibodies, the primary antibodies or their IgG controls were diluted in 200 $\mu$l of blocking solution in Eppendorf tubes and the tubules were then transferred into them. The tubules were incubated for 1 hour and 30 minutes on rotating wheel at room temperature. They were then moved back to a 24-well plate and washed 3 times for 15 minutes with blocking solution with agitation. The tubules were incubated with DAPI diluted 1:1000 in TBS for 10 minutes, placed in a clean microscope slide. A few drops of VectaShield® Antifade Mounting Medium was added and covered with a glass coverslip without air bubbles. To compare possible changes in specificity of an antibody after fluorophore conjugation, the tubules were first incubated with the unconjugated primary antibody or the IgG control as described above. They were then incubated with the secondary antibodies in 500 $\mu$l of blocking solution per tube for 2 hours at room temperature in agitation. The tubules were washed 3 times for 15 minutes with 2%BSA, 0.1% Triton X-100 in TBS in agitation. Then, they were incubated with the corresponding fluorophore conjugated primary antibodies or corresponding IgG controls in 200$\mu$l of blocking solution in 200 $\mu$l eppendorf tubes for 1.5 hours on rotating wheel at room temperature. The tubules were washed as described before, incubated with DAPI, placed on microscope slide, added with VectaShield® Antifade Mounting Medium and covered with a glass as described before. The list of antibodies and dilutions is included in the Appendix.

## 2.20 FACS staining of testicular single cell suspension

Cells were incubated with 1 $\mu$l Near-IR fluorescent reactive dye LIVE/DEAD™ Fixable Dead Cell Stain Kit (Invitrogen; Cat: L34961; Ex: 633/635nm; Em:775nm) per tube of 1 ml on ice for 30 minutes. 500 $\mu$l of HBSS + 5% FBS was added, and the cells centrifuged at 2000g for 5 minutes at 10℃. The supernatant was discarded, and the cells were blocked with 1 ml of HBSS + 5% FBS on ice for 45 minutes. They were again centrifuged at 2000g for 5 minutes at 10℃ and the supernatant discarded. They were resuspended in 150 $\mu$l of Fix and Perm Solution A per 1 million cells and incubated at room temperature for 30 minutes. After adding 600 $\mu$l of HBSS + 5% FBS, they were centrifuged at 2000g for 10 minutes at 10℃ and the supernatant discarded. They were resuspended in 150 $\mu$l of Fix and Perm Solution B per 1 million cells and incubated at room temperature for 30 minutes. 500 $\mu$l of HBSS + 5% FBS was added, and the cells centrifuged at 2000g for 10 minutes at 10℃. The cells were then incubated with primary antibodies or IgG controls in HBSS +5% FBS for 1h at RT. They were

centrifuged at 2000g for 10 minutes at 10℃ and the supernatant discarded. They were resuspended in 200 $\mu$l of HBSS + 5% FBS and 2 drops of Hoechst (NucBlue® Live ReadyProbes® Reagent, R37605, Thermo Fisher) was added per tube of 100 $\mu$l, vortexed briefly and incubated at 4℃ for 20 minutes in dark without shaking. The cells were then sorted in BD FACS Aria Fusion. The list of antibodies and dilutions is included in the Appendix. Sorted populations used for Micro-C were stored at 4℃ overnight.

## 2.21 Low input Micro-C

### 2.21.1 Crosslinking

The FACS sorted cells were processed for Micro-C library preparation. The cells were pelleted by centrifuging for 5 min at 1000g at room temperature and washed with 1X PBS. Since they were already cross-linked with formaldehyde for the FACS sorting, the pellets were carried forward to the second cross-linking step of Micro-C using the long cross-linkers EGS (Sigma) at a final concentration of 3 mM in PBS for 45 min at room temperature with passive mixing. The reaction was quenched with 0.4% 1M Glycine for 5 min at room temperature. Simultaneously, BSA was added to a final concentration of 1% to prevent cell sticking. The cell pellet was washed once with 1X PBS and taken forward for Micro-C library preparation.

### 2.21.2 Preparation of Concanavalin A-coated beads

Concanavalin A-coated bead slurry was taken at 10 $\mu$l per sample and washed twice with 1X Binding buffer (20 mM HEPES-KOH pH 8, 10 mM KCl, 1 mM CaCl2, 1 mM MnCl2, 0.5 mM spermidine) using a magnet stand. Then, it was resuspended in 1X Binding buffer equivalent to 10 $\mu$l per sample.

### 2.21.3 Chromatin digestion by MNase

Micro-C libraries were constructed according to Hsieh, Cattoglio, et al., 2020 with modifications optimized in our lab to process samples with low cell numbers down to 10,000 cells.

Intact nuclei were extracted by treating cells with Micro-C Buffer 1 (MB1) (50 mM NaCl, 10 mM Tris-HCl pH = 7.5, 5 mM MgCl2, 1M CaCl2, 0.2% NP-40, 1x Protease Inhibitor Cocktail (Sigma 5056489001)) for 20 min on ice. The nuclei were centrifuged for 5 min at 5000g at 4℃. The supernatant was discarded and the nuclei resuspended in MB1 by gently pipetting. 10X binding buffer (200 mM HEPES-KOH pH 8, 100 mM KCl, 10 mM CaCl2, 10 mM MnCl2, 5 mM spermidine) was added to the nuclei solution to a final concentration of 1X. 10 $\mu$l bead slurry was added per sample and incubated with rotation for 10 min at room temperature. The nuclei on beads were washed in MB1 and resuspended in 500 $\mu$l MB1. Chromatin was digested with a pre-titrated MNase concentration (Worthington Biochem LS004798) at 37℃ for 10 min with shaking at 950 rpm. MNase digestion was stopped by adding 4 mM EGTA. Digested chromatin was washed twice with ice-cold Micro-C Buffer 2 (50 mM NaCl, 10 mM Tris-HCl pH = 7.5, 10 mM MgCl2) on a magnet stand.

### 2.21.4 End repairing and labeling

Chromatin was incubated with $\sim$ 10U/1$\mu$g DNA T4 PNK in Micro-C end-repair buffer (50 mM NaCl, 10 mM TrisHCl pH 7.5, 10 mM MgCl2, 100 $\mu$g/mL BSA (New England BioLabs B9000), 2 mM ATP (Thermo Fisher Scientific R1441), 5 mM DTT) at 37℃ for 15 min. 10U/1$\mu$g DNA Klenow Fragment was added and incubated at 37℃ for 15 min. The blunting and labeling reaction was triggered upon adding biotin-dATP (Jena Bioscience NU-835-BIO14), biotin-dCTP (Jena Bioscience NU-809-BIOX), dGTP, and dTTP to a final concentration of 66 mM each. Additionally, BSA at final concentration of 100 $\mu$g/mL and T4 DNA Ligase Buffer was added and the solution incubated for 4 hours at 25℃. At 2 hours mid-point, d 1.5 $\mu$L 100 mM ATP and 1 $\mu$L 50 U Klenow was added. T4 PNK and Klenow Fragment were inactivated by adding 30 mM EDTA. Biotin-labeled chromatin was washed twice by ice-cold Micro-C Buffer 3 (50 mM TrisHCl pH 7.5, 10 mM MgCl2).

### 2.21.5 Proximity ligation and removal of biotin-dNTP from un-ligated ends

Crosslinked nucleosomes were ligated by T4 DNA Ligase (New England BioLabs M0202) 250 $\mu$L solution with BSA at final concentration of 100 $\mu$g/mL at 22 ℃ with slow rotation for at least 4 hours. 2.5 uL 2000 U/$\mu$l T4 DNA ligase was added again at 2 hour mid-point. After placing on the magnet stand to separate the beads from the supernatant, the supernatant was discarded and the biotin-dNTPs on un-ligated ends were removed using Exonuclease III (New England BioLabs M0206) and incubating at 37℃ for 15 min with interval mixing. The chromatin was reverse crosslinked by incubating with 2mg/mL Proteinase K and 1% SDS overnight at 65℃.

### 2.21.6 DNA purification and Library Preparation

The DNA was purified with phenol-chloroform extraction and resuspended in TE buffer (10 mM Tris HCl pH 8.0, 1 mM EDTA) with 1X RNase A. It was quantified by Qubit. The purified DNA with biotin-dNTPs was captured by Dynabeads® MyOne Streptavidin C1 (Thermo Fisher Scientific 65001). Standard Illumina library preparation protocol including end-repair, A-tailing, and adaptor ligation was performed on beads with the NEBnext Ultra II kit (New England BioLabs E7645). An optimal PCR cycle for final library amplification was determined by quantification PCR (KAPA Biosystems KK4602). I performed 11-14 cycles for the sorted populations of testicular cells. The sequencing library was amplified by Kapa HiFi PCR enzyme (KAPA Biosystems KK2601) with the lowest possible cycles to reduce PCR duplicates. Two-sided size selection was performed using AMPure XP beads to only keep fragments in the 370-620 bp range (250-400 di-nucleosomal ligation products plus 120 bp adaptors). The library concentration and fragment size distribution were measured by Qubit and Bioanalyzer respectively. For the shallow sequencing dataset analysed in this thesis, the libraries was sequenced by paired-end sequencing with 150 bp to a depth of 1 million read pairs per library in Illumina MiSeq sequencer.

## 2.22 Micro-C analysis

Micro-C data were analyzed using FAN-C v0.9.25 (Kruse et al., 2020). Paired-end reads were scanned to identify ligation junctions, split at ligation junctions, if any were present, and then aligned independently to the GRCh38.p13 genome using BWA-MEM (version 0.7.17-r1188) (Li and Durbin, 2009). Aligned reads were filtered to retain only uniquely aligned reads with a mapping quality of at least three. Reads were then paired based on read names and assigned to 100-bp genomic bins. 'Inward' and 'outward' reads from the same or adjacent bins (separated by less than 50 bp) were removed. In addition, we removed PCR duplicates, reads mapping more than 500 bp from the genomic bin and self-ligations, for which both reads map to the same genomic bin. Matrices were created, binned to a resolution of 2 Mb and filtered using FAN-C default parameters to remove bins with coverage less than 10% of the median coverage. Normalization was performed using Knight-Ruiz matrix balancing (Knight and Ruiz, 2013). PCA analysis was performed for the 2Mb resolution matrices using FANC with default parameters and after ignoring pixels with no contacts in any sample.

I generated three biological replicate datasets for each developmental stage, which showed high similarity in PCA. Therefore, I pooled the matrices of the biological replicates. Matrices created from merged biological replicates were again binned to a resolution of 2 Mb and filtered and normalized as before. Expected contacts were calculated as the average contacts at each genomic distance separation. Hi-C data were visualized using plotting tools from FAN-C.

# Chapter 3

# Generation and characterization of joint snATAC-seq + snRNA-seq multiome datasets of the adult human testis

## 3.1 Introduction

Spermatogenesis describes the process of production of the male gametes, and it occurs within the seminiferous tubules of the testis. The process involves a continuum of germ cell development where spermatogonial stem cells undergo sequential mitosis, meiosis and differentiation to produce the haploid spermatids. Germ cell development is further supported by somatic cells that reside within and outside the seminiferous tubule.

Apart from housing a diverse composition of cell types, the testis also has one of the most complex transcriptomes in the human body (Xia et al., 2020; Soumillon et al., 2013). The inherent heterogeneity of the tissue, coupled with the small size of biopsy samples obtained in humans, makes it challenging to study the testis through bulk -omics approaches. The advent of single cell technology has allowed the characterization of the human testis at high resolution across development and ageing (Guo, Nie, et al., 2020; Guo, Sosa, et al., 2021; Garcia-Alonso et al., 2022; Nie et al., 2022; Sohni et al., 2019; Zhao et al., 2020), during altered spermatogenesis (Guo, Nie, et al., 2020; Di Persio, Tekath, et al., 2021; Wang, Liu, et al., 2018; Zhao et al., 2020) and between multiple species (Murat et al., 2023; Shami et al., 2020).

Despite extensive studies mapping spermatogenesis in various biological contexts, a comprehensive understanding of the key transcription factors and regulatory elements that drive normal human spermatogenesis is lacking. This requires the joint single-cell assessment of both chromatin state and gene expression in order to robustly infer their relationships and regulatory dynamics. While the adult testis has previously been investigated using multiomic approaches (Wu et al., 2022), it involved the integration of snATAC-seq and scRNA-seq datasets that were generated independently from different cells. While Garcia-Alonso et al., 2022 have indeed produced

snATAC-seq and snRNA-seq datasets from the same single cells, they investigate the fetal gonad and thus do not capture the gene regulatory networks governing mature sperm development which only begins at puberty.

In this chapter, we have generated the first snATAC-seq + snRNA-seq multiome libraries from the same single cells of the adult human testis. Upon optimization of nuclei isolation for library generation, we capture cell type-specific gene expression and accessibility patterns in germ cells and supporting somatic cells throughout spermatogenesis. In comparison with published scRNA-seq datasets, we capture more spermatogonia and fewer spermatids due to the technicality of the nuclei isolation procedure as explained further in the chapter. Finally, all clusters in our data show high correlation of the gene expression patterns with the corresponding cell types in the published datasets. Hence, our data can be used to generate a comprehensive single nucleus map of chromatin accessibility, gene expression and gene regulation during healthy and dysfunctional human spermatogenesis, which can be used to inform the interpretation of genes and genomic variation implicated in male infertility (see Appendix).

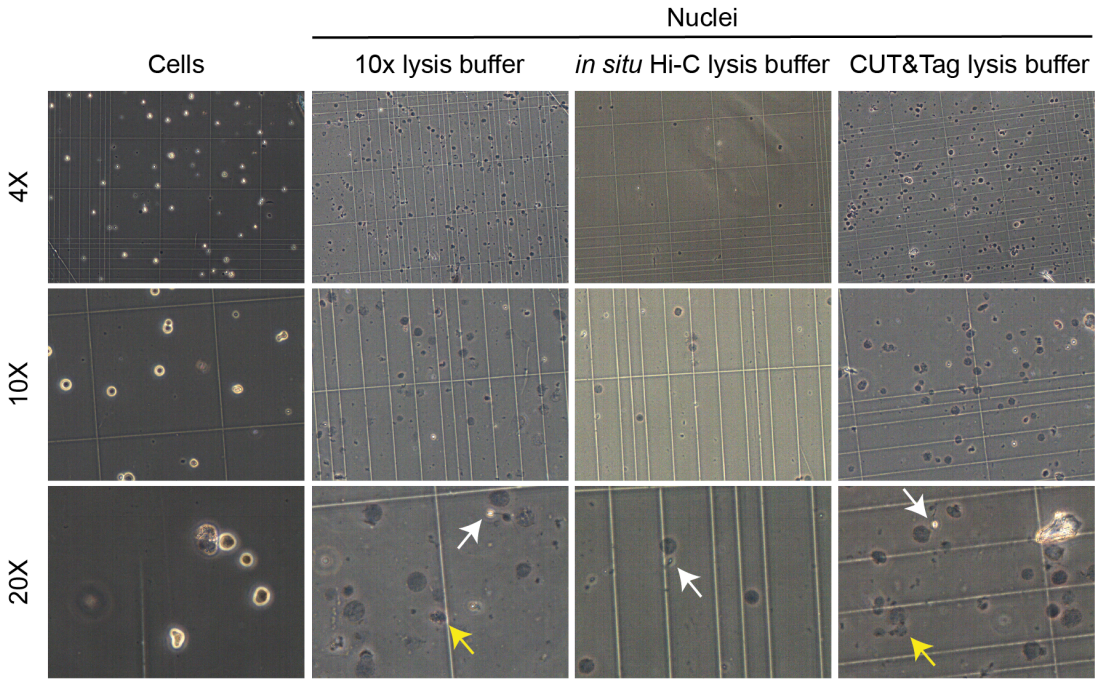
## 3.2 Optimization of nuclei isolation from the adult human testis

The adult human testis is a heterogeneous tissue consisting of germ cells and supporting somatic cells of various sizes and DNA content. To simultaneously capture the transcriptome and accessible regions of the genome from the same single cells, we first prepared single cell suspensions from tissue biopsies using a two-step procedure of enzymatic digestion and physical filtering (Kossack et al., 2013; Laurentino et al., 2019; Neuhaus et al., 2017). I then optimized nuclei isolation from the single cells to use as input for the 10x multiome snATAC + Gene Expression platform.

For nuclei isolation, I adapted “Low Cell Input Nuclei Isolation” from 10x Genomics Nuclei Isolation for Single Cell Multiome ATAC + Gene Expression Sequencing Revision A (CG000365) to include minor changes in centrifugation speed and time (see Methods) and major changes in buffer composition and lysis timings (discussed below). All steps were performed at 4℃. I estimated optimal lysis conditions based on nuclear integrity when visualized at 20X magnification in a light microscope. When stained with Trypan blue, intact cells appear spherical with a bright ring around the periphery and nuclei appear as dark flat circular discs. Over-lysed nuclei show irregular nuclear membrane and blebbing. I also counted the number of high quality nuclei recovered in the Neubauer Haemocytometer Counting Chamber. Use of an automated cell counter like LUNA-FL™ Dual Fluorescence Cell Counter was inaccurate in estimating the number of nuclei due to the incorrect inclusion of debris in nuclear count (not shown).

Firstly, I tested different compositions of the lysis buffer. I adapted the following lysis buffers from other protocols and incubated the cells for 40 seconds at 4℃ - 10x lysis buffer (10 mM Tris HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% Tween-20, 0.1% NP40, 0.01% Digitonin, 1% BSA, 1mM DTT), adapted from Nuclei Isolation for Single Cell Multiome ATAC + Gene Expression Sequencing Revision A (CG000365);

A




B

| Lysis Buffer | Number of Cells | Number of Nuclei | Recovery (%) |
|---|---|---|---|
| 10x | 200,000 | 13,500 | 6.75 |
| *in situ* Hi-C | 200,000 | 2,300 | 1.15 |
| CUT&Tag | 200,000 | 21,000 | 10.5 |

Figure 3.1: ***Optimization of lysis buffer composition for nuclei isolation from adult human testicular cells.*** **(A)** Light microscopy images of adult human testicular cells or nuclei incubated for 40 seconds at 4℃ with 10x, *in situ* Hi-C or CUT&Tag lysis buffers at 4X, 10X and 20X magnification. Unlysed cells and blebbing nuclei in the nuclear suspension are indicated in white and yellow arrows respectively. **(B)** Table of counts for nuclei recovered under each of the lysis conditions. The cells and nuclei were counted in the Neubauer Haemocytometer Counting Chamber.

CUT&Tag lysis buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1X Protease Inhibitor Cocktail, 0.01% NP40, 0.01% Digitonin, 1mM DTT), adapted from Kaya-Okur et al., 2019, and *in situ* Hi-C lysis buffer (10 mM Tris pH 8.0, 10 mM NaCl, 0.2% IGEPAL-360 and 1X Protease Inhibitor Cocktail) adapted from Díaz et al., 2018. I added 1 U/$\mu$l RNase inhibitor to the lysis buffer during final multiome library preparation (see Methods).

CUT&Tag lysis buffer showed maximum recovery of high quality nuclei (10.5 %) followed by 10x lysis buffer (6.75%) and *in situ* Hi-C lysis buffer (1.15%) (Figure 3.1). Hence, I decided to proceed with CUT&Tag lysis buffer for rest of the protocol.

Next, I optimized the timing of lysis buffer incubation. 10x Genomics recommends 3 minutes for cryopreserved Peripheral Blood Mononuclear Cells (PBMCs) and 5 minutes for cryopreserved cell lines (CG000365, Revision A). To compare the sensitivity of primary human testicular cells to more commonly used cell lines, I also isolated nuclei from HEK-293T (Human Embryonic Kidney) cells using the CUT&Tag lysis buffer. Since HEK-293T nuclear suspension presented unlysed cells even after 10 minutes of incubation with CUT&Tag lysis buffer (white arrow, Figure 3.2), I decided to incubate them for 13 minutes in the final comparison.

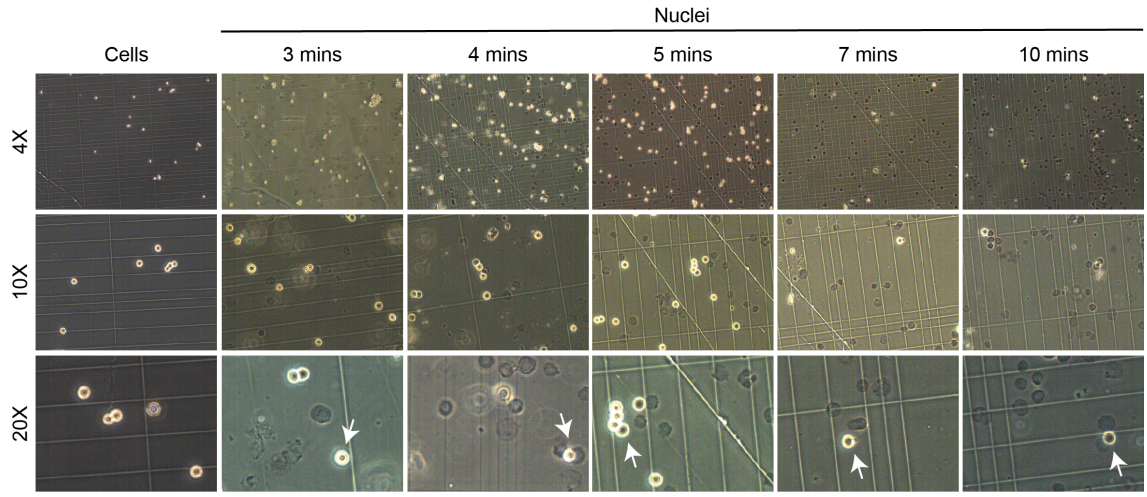


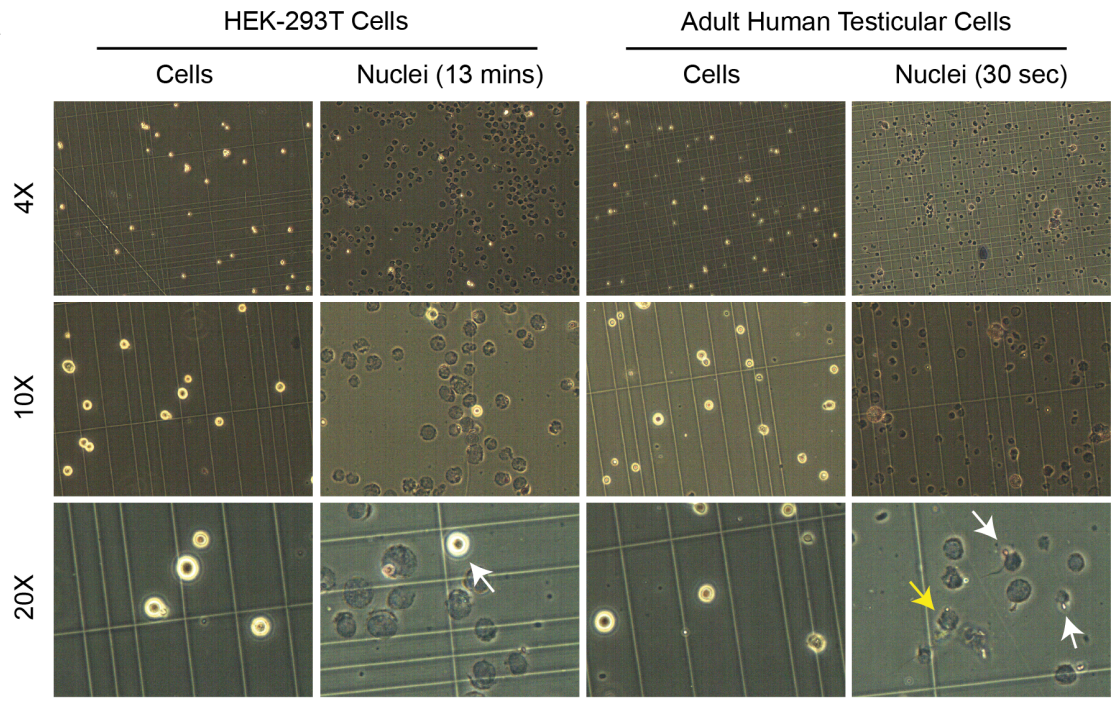

Figure 3.2: ***Optimization of lysis timing for nuclei isolation from HEK-293T cells.*** Light microscopy images of HEK-293T cells or nuclei incubated for 3, 4, 5, 7 and 10 minutes at 4℃ with CUT&Tag lysis buffer at 4X, 10X and 20X magnification. Unlysed cells in the nuclear suspension are indicated with white arrows.

As seen in Figure 3.3, incubation of HEK-293T cells for 13 minutes at 4℃ in CUT&Tag lysis buffer results in a majority of successfully lysed cells. However, the adult human testicular cells can only be incubated in the same lysis buffer for ∼30 seconds. In the testicular nuclear suspension, while some nuclei start showing signs of blebbing (yellow arrows), there are still some unlysed elongating spermatids (white arrows) (Figure 3.3A). This indicates that an incubation time of 30 seconds is too long for some cells and too short for some others, highlighting the heterogeneity of the testicular suspension and the associated challenge of finding a common lysis condition that is optimal for all cells in the tissue. Furthermore, HEK-293T cells show 60% recovery while the latter only show 30% recovery (Figure 3.3B). Hence, adult human testicular cells are extremely sensitive compared to commonly used cell lines and require lysis period in the range of a few seconds at 4℃.

To further optimize a time range for efficient recovery of nuclei from all cells in the testicular suspension, I incubated the cells with CUT&Tag lysis buffer for 20, 30 and 40 seconds at 4℃. I still observed unlysed elongating spermatid heads (white arrows) in all three conditions (Figure 3.4A), with a marked reduction in recovery rate and an increase in over lysed nuclei (yellow arrows) at 40 seconds (Figure 3.4B).

To get a more global view of nuclei composition, I stained fresh cells and nuclei recovered after 30-second incubation in CUT&Tag lysis buffer with Hoechst 33342 DNA stain and observed the signal in flow cytometry. The testicular cells vary in their DNA content. Compared to spermatogonia (or somatic cells and secondary spermatocytes), the primary spermatocytes carry double the amount of DNA and the haploid spermatids carry half the amount of DNA. This results in a peak of 2C, 4C and 1C for spermatogonia/secondary spermatocytes/somatic cells, primary spermatocytes and haploid spermatids respectively. Within the spermatid population, late elongating spermatids have more condensed DNA than early round spermatids. Thus, they absorb less Hoescht stain and, occasionally, appears as a separate HC (hypohaploid) peak with lower intensity than 1C peak. In Figure 3.4C, we can see a depletion of HC peak in nuclei compared to intact cells, in line with our microscopy results of unlysed

A




B

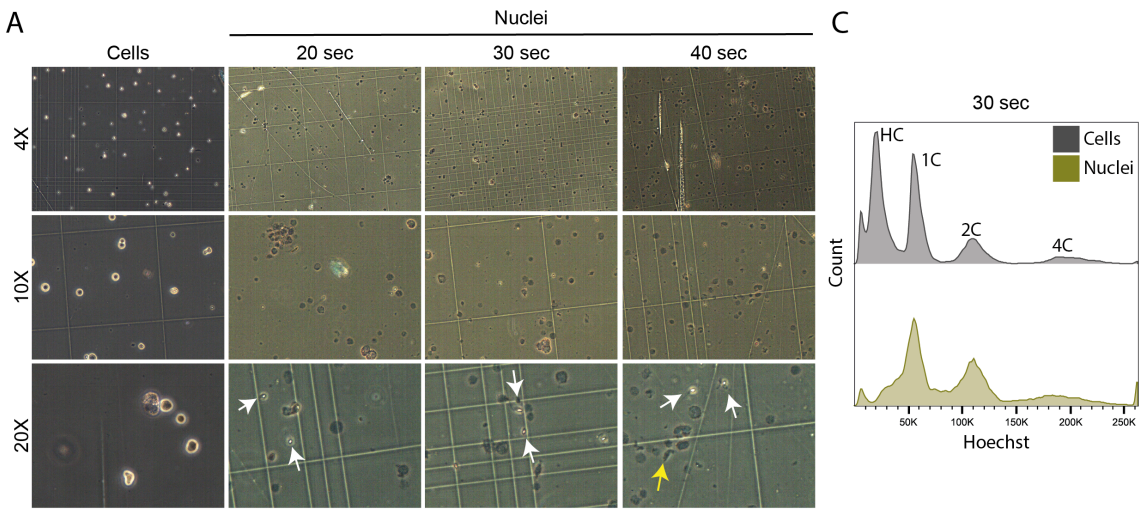
| Cell Type | Number of Cells | Number of Nuclei | Recovery (%) |
|---|---|---|---|
| HEK-293T Cells | 1,000,000 | 600,000 | 60 |
| Adult Human Testicular Cells | 2,000,000 | 615,000 | 30.75 |

Figure 3.3: ***Optimization of lysis time for nuclei isolation from HEK-293T cells and adult human testicular cells.*** **(A)** Light microscopy images of HEK-293T cells or adult human testicular cells or nuclei obtained from incubation with CUT&Tag lysis buffer for 13 minutes or 30 seconds at 4℃ respectively at 4X, 10X and 20X magnification. Unlysed cells and blebbing nuclei in the nuclear suspension are indicated in white and yellow arrows respectively. **(B)** Table of counts for nuclei recovered under each of the lysis conditions. The cells and nuclei were counted in the Neubauer Haemocytometer Counting Chamber.

elongating spermatid heads in the nuclear suspension.

In an attempt to bring all cells to the same level of sensitivity and expose them to the same lysis conditions, I fixed the single cell suspension of the adult human testis with 80% Methanol (MeOH) for 7 minutes at -20℃. At the time of optimization, 10x Genomics had released a successful protocol to produce scRNA-seq libraries from MeOH-fixed cells (CG000136, Revision E). Upon enquiry, the technical support team also mentioned seeing promising results in pilot experiments attempting to generate snATAC-seq libraries from MeOH-fixed PBMCs.

In addition to its fixing activity, MeOH also permeabilizes cells, making them appear blue when stained with Trypan blue. I could not obtain nuclei from MeOH-fixed cells upon lysis with CUT&Tag lysis buffer even after 5 or 15 minutes of incubation at 4℃ (Figure 3.5A). Hence, I repeated the isolation process with 10x lysis buffer, a composition that proved to be harsher on fresh adult human testicular cells in earlier experiments (Figure 3.1). Again, I did not obtain any nuclei. Hence, nuclei could not be isolated from MeOH-fixed adult human testicular cells with CUT&Tag or 10x lysis buffers.




B

| Time (s) | Number of Cells | Number of Nuclei | Recovery (%) |
|---|---|---|---|
| 20 sec | 200,000 | 27,000 | 13.5 |
| 30 sec | 200,000 | 26,400 | 13.2 |
| 40 sec | 200,000 | 21,100 | 10.55 |

Figure 3.4: ***Optimization of lysis timing for nuclei isolation from adult human testicular cells.*** **(A)** Light microscopy images of adult human testicular cell suspension or nuclei incubated for 20, 30 or 40 seconds at 4℃ with CUT&Tag lysis buffer at 4X, 10X and 20X magnification. Unlysed cells and blebbing nuclei in the nuclear suspension are indicated with white and yellow arrows respectively. **(B)** Table of counts for nuclei recovered under each of the lysis conditions. The cells and nuclei were counted in the Neubauer Haemocytometer Counting Chamber. **(C)** FACS plots of adult human testicular cells or nuclei isolated with 30-second incubation in CUT&Tag lysis buffer. Cells and nuclei were stained with Hoechst 33342.

To circumvent this challenge, I attempted to use MeOH-fixed cells, instead of nuclei, as input for multiome library generation. As observed in pilot experiments of 10x Genomics, I reasoned that the permeabilization achieved during MeOH fixation could allow successful entry of ATAC-seq transposition reagents into the nucleus of cells. As control, I generated multiome libraries using nuclei isolated from fresh cells upon incubation in CUT&Tag lysis buffer for 30 seconds at 4℃.

snRNA-seq libraries generated using MeOH-fixed cells or nuclei from fresh cells exhibit the expected peak at ∼500bp in its fragment size distribution (Figure 3.5B, top panel). However, snATAC-seq libraries show the typical nucleosomal patterning with peaks at ∼250bp for mono-, ∼400bp for di- and ∼650bp for tri-nucleosomes only when produced using nuclei isolated from fresh cells and not when produced from MeOH-fixed cells (Figure 3.5B, bottom panel).

Hence, I proceeded to use nuclei isolated from fresh adult human testicular cells after incubation in CUT&Tag lysis buffer for ∼30 seconds at 4℃ as input for all the snATAC-seq + snRNA-seq multiome libraries generated and analysed in this project.

**Contributions** As part of the snATAC-seq + snRNA-seq multiome library generation of the adult human testis, we received testicular biopsies of adult humans presenting obstructed vas deferens or undergoing vasectomy reversal from Centrum für Reproduktionsmedizin und Andrologie (CeRA) in Münster, Germany. Sara Di Persio (CeRA) or Nicole Tewort (CeRA) performed enzymatic digestion of the tissue to generate sin-

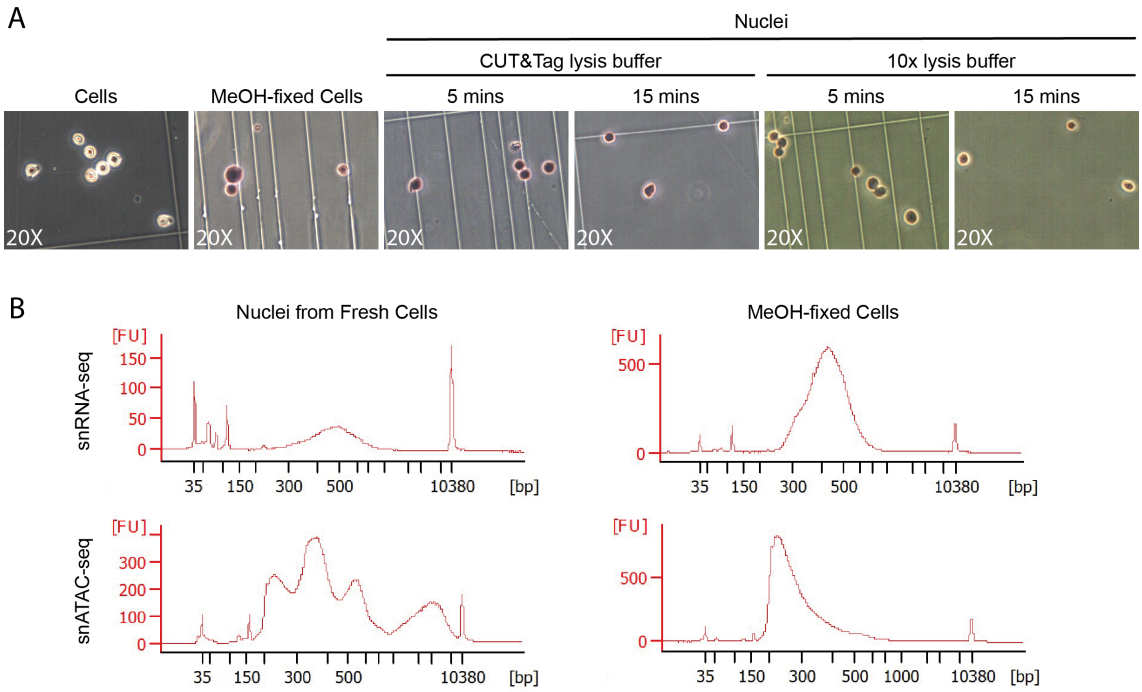



Figure 3.5: ***10x snATAC + snRNA multiomic library generation from MeOH-fixed adult human testicular cells.*** **(A)** Light microscopy images of fresh or MeOH-fixed adult human testicular cells and nuclei isolated from MeOH-fixed cells with CUT&Tag or 10x lysis buffer incubated for 5 and 15 minutes at 4℃ each. **(B)** Bioanalyzer tracks of 1 $\mu$l of snRNA-seq or snATAC-seq libraries generated from MeOH-fixed cells or nuclei isolated from fresh cells.

gle cell suspensions. I designed and performed the optimization of nuclei isolation from these single cell suspensions. I also optimized the generation of multiome libraries and generated two out of three libraries from healthy samples analysed in this chapter while Sara Di Persio generated the third replicate.

## 3.3 snATAC-seq + snRNA-seq multiome captures all known cell types of the adult human testis

Using nuclei isolated from fresh single cell suspension, we generated single nucleus multiome data (snATAC + snRNA) from testis biopsies from three men with healthy sperm development according to clinical parameters (see Appendix).

After I performed standard quality control and filtering (see Methods), > 19,000 high quality cells were retained for downstream analysis, with a mean of $\sim$ 2,520 expressed genes and a mean of $\sim$2,840 ATAC-seq peaks detected per cell. snRNA-seq and snATAC-seq data were separately processed and then integrated using a weighted nearest-neighbour (WNN) approach.

Based on known marker gene expression (Figure 3.6C), we identified 14 germline clusters and seven somatic clusters (Figure 3.6B). Clusters 4, 6, 11, 2, 14, 5, 16, 1, 13, 8, 15, 17, 12 and 10 correspond to germ cells (germ cell marker *DDX4*) and cluster 0, 3, 7, 18, 20, 22 and 21 correspond to somatic cells (somatic cell marker *VIM*). The germ cells and somatic cells cluster separately, and the germ cells arrange

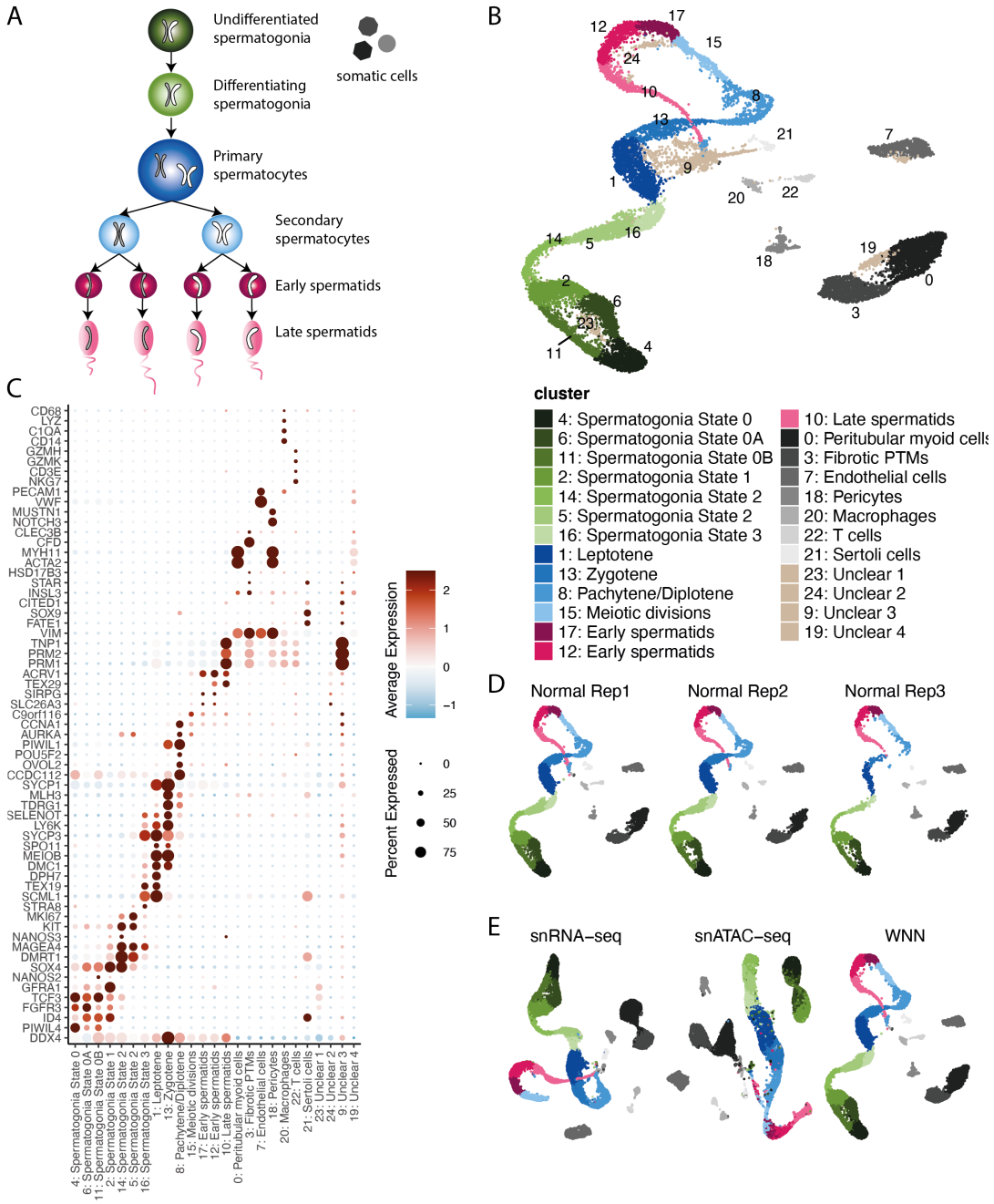



Figure 3.6: ***snATAC-seq + snRNA-seq multiomic characterization of the adult human testis.*** **(A)** Schematic overview of the stages of human spermatogenesis. Undifferentiated spermatogonia commit to differentiation, then become spermatocytes which go through meiosis to produce haploid spermatids. The seminiferous tubules also contain Sertoli cells, and the interstitial space between tubules contains Leydig cells, peritubular myoid cells (PTMs), pericytes, endothelial cells, and immune cells. **(B)** WNN UMAP representation of single-nucleus multiomics data from adult human testis. "Unclear" clusters could not be identified based on marker gene expression and were excluded from further analysis. **(C)** Marker gene expression across clusters used to assign cluster identity. Colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression. **(D)** WNN UMAPs of the three biological replicates shown separately. **(E)** UMAPs based on snRNA-seq, snATAC-seq, and the weighted nearest neighbour (WNN) graph produced by combining the two modalities using Seurat.

themselves in the sequence of developmental order.

Within the germ cell compartment, clusters 4, 6, 11 and 2 correspond to undifferentiated spermatogonia (*ID4*). Clusters 4, 6 and 11 express *PIWIL4* which was identified as a marker for the most naive State 0 stem cell population in Guo, Grow, Mlcochova, et al., 2018. Clusters 6 and 11 correspond to State 0A (*FGFR3*) and State 0B (*NANOS2*), two subgroups of State 0 identified in Di Persio, Tekath, et al., 2021; Sohni et al., 2019. Cluster 2 is State 1 spermatogonia (*GFRA1*). Clusters 14, 5 are differentiating spermatogonia State 2 (*KIT, DMRT1, MKI67*) and cluster 16 is differentiating spermatogonia State 3 (*STRA8*). Cluster 1, 13, 8 and 15 correspond to meiotic spermatocytes where cluster 1 is leptotene (*SCML1, TEX19, DPH7, DMC1, MEIOB, SPO11, SYCP3*), cluster 13 is zygotene (*LY6K, SELENOT, TDRG1, MLH3, MEIOB, SPO11, SYCP1*), cluster 8 is pachytene/diplotene (pachytene markers: *CCDC112, OVOL2, POU5F2, PIWIL1*; diplotene markers: *AURKA, CCNA1*) and cluster 15 is meiotic divisions (*C9ORF116, SLC26A3, SIRPG*). Cluster 15 has low expression for *C9ORF116* and almost no expression for *SLC26A3, SIRPG*. However, it exhibits bimodal distribution of ATAC-seq reads from chrX, indicating the separation of cells into chrX or chrY-bearing cells after meiosis I (Chapter: Comparison of -X and -Y spermatids; Figure 5.2). Thus, we have assigned cluster 15 as secondary spermatocytes that undergo meiosis II divisions. Clusters 17 and 12 correspond to early spermatids (*TEX29, ACRV1*) and cluster 10 corresponds to late spermatids (*PRM1, PRM2, TNP1*).

Amongst the somatic cell clusters, cluster 0 is peritubular myoid cells (*ACTA2, MYH11*), cluster 3 is fibrotic peritubular myoid cells (Fibrotic PTMs) (*CFD, CLEC3B*) (Di Persio, Tekath, et al., 2021), cluster 7 is endothelial cells (*VWF, PECAM1*), cluster 18 is pericytes (*MUSTN1, NOTCH3*), cluster 20 is macrophages (*CD14, C1QA, LYZ, CD68*), cluster 22 is T cells (*NKG7, CD3E, GZMK, GZMH*) and cluster 21 is Sertoli cells (*FATE1, SOX9, CITED1*). Lastly, clusters 23, 24, 9 and 19 could not be assigned to a known cell type as they show mixed germline-somatic expression. These "Unclear" clusters were excluded from further analysis.

We do not identify a cluster that uniquely corresponds to Leydig cells (*INSL3, STAR, HSD17B3*) in our dataset. Further examination revealed that we capture very few Leydig cells that are distributed amongst the clusters corresponding to peritubular myoid cells, Fibrotic PTMs, Sertoli cells and the unclear cluster assignment (data not shown). The limited number of Leydig cells captured could be because they exist in the interstitial compartment between seminiferous tubules. During enzymatic digestion, we first generate a suspension of individual tubules, and then we digest it to produce the single cell suspension. Thus, we probably enrich for cells that exist within the tubule compared those that reside outside it. Additionally, similar to Guo, Grow, Mlcochova, et al., 2018; Sohni et al., 2019, we also obtain very few Sertoli cells. This has been attributed to their size and cellular architecture, making it difficult to be captured during the single cell encapsulation process. The elongated shape and large size of the cells could also respectively make them vulnerable to breakage and prevent their passage through the filter during the single cell isolation procedure employed in this study.

Figure 3.6D depicts high similarity in the UMAPs of the individual biological replicates and Figure 3.6E depicts UMAPs obtained from snRNA-seq alone, snATAC-seq alone or WNN. We can see the continuum of germ cell clusters along developmental trajectory and the separation of germ cells and somatic cells from each other in all

three cases. However, the germ cell continuum is disconnected before 15: Meiotic divisions in snRNA-seq UMAP and before the first differentiating spermatogonia cluster of 14: Spermatogonia State 2 in snATAC-seq UMAP. This could indicate a significant developmental transition in the transcriptome from meiosis I to meiosis II and in the chromatin accessibility landscape from undifferentiated to differentiating spermatogonia.

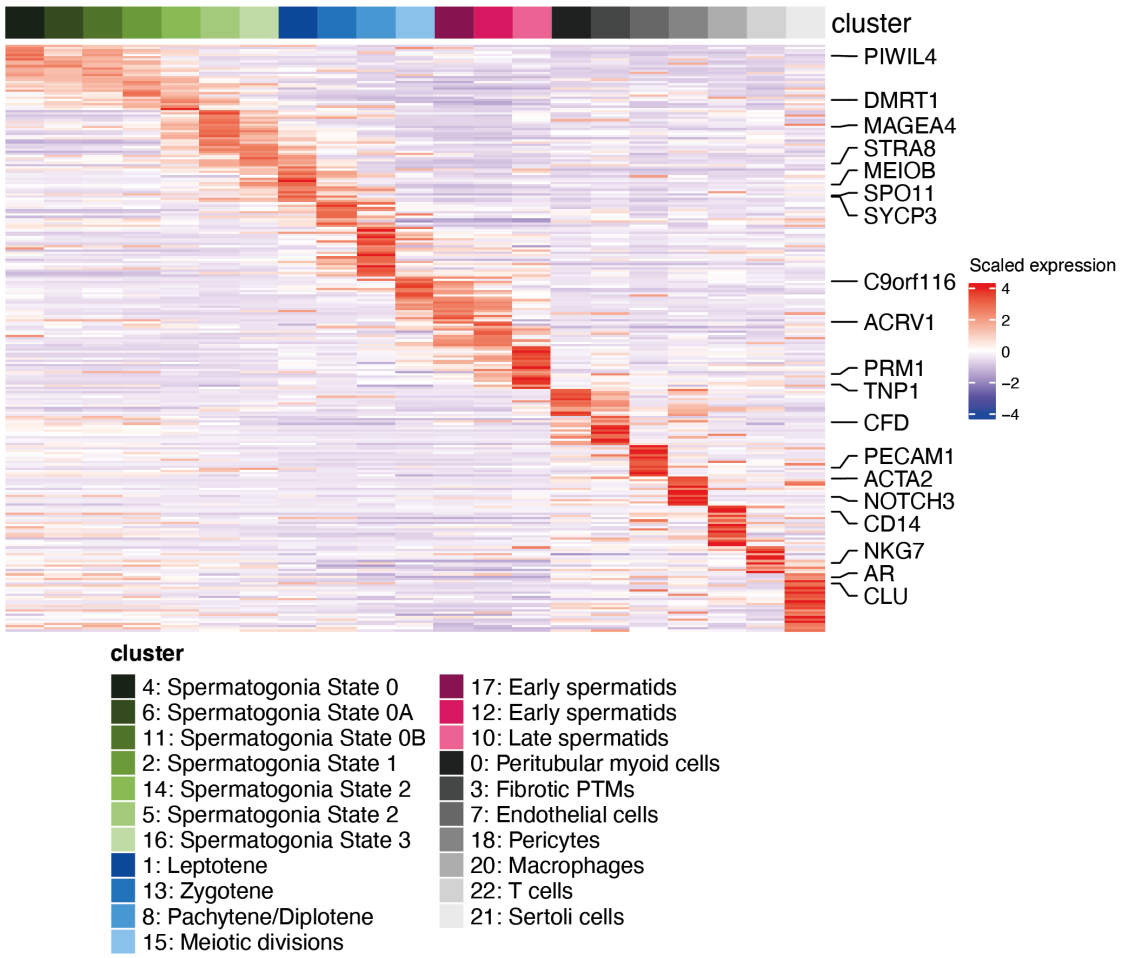



Figure 3.7: ***Heatmap of gene expression across adult human spermatogenesis.*** Clusters are ordered by the developmental trajectory, followed by somatic cells. Genes are ordered by the cluster in which the gene has max expression. Selected marker genes for different cell types are labelled on the right-hand side.

Differential gene expression analysis identified a total of ∼12,000 genes displaying cluster-specific expression, including many known markers for the different cell types in the testis that are highlighted in the heatmap (Figure 3.7).

Examining chromatin accessibility at these loci shows dynamic promoter and gene body accessibility which correlates with gene expression. Figure 3.8 depicts example loci of UTF1, an undifferentiated spermatogonia marker, TEX19, a leptotene marker and TNP1, a spermatid marker. In addition, we can also identify distal peaks with stage-specific accessibility during spermatogenesis (grey arrow in third panel "TNP1"), which represent putative enhancers.

In summary, our dataset captures the whole known repertoire of spermatogenesis including germ cells at different stages of development and most supporting somatic

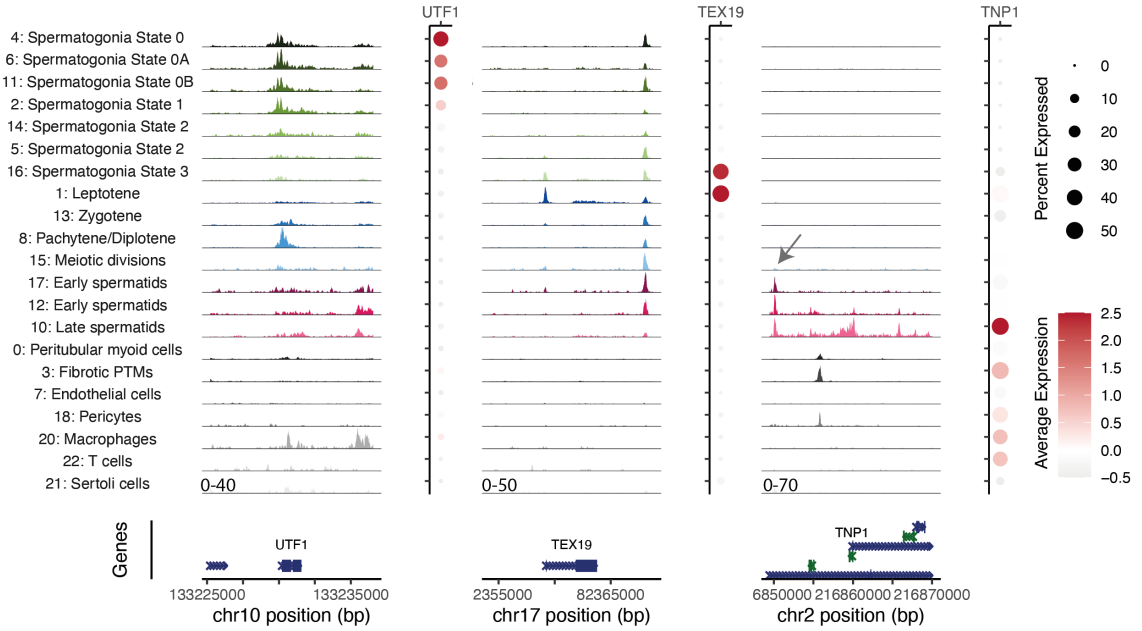



Figure 3.8: ***Examples of dynamic chromatin accessibility at selected marker genes.*** Each subpanel shows cluster-specific pseudobulk chromatin accessibility tracks on the left. Expression of the marker gene of interest in the cluster is shown in the dot plots on the right, where colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression.

cells in the testis. Furthermore, joint analysis of snATAC-seq and snRNA-seq of the adult human testis helps resolve the identity of short-lived cell types like secondary spermatocytes, that do not show very high expression of most marker genes.

## 3.4 snRNA-seq captures more spermatogonia and fewer spermatids than scRNA-seq

While the 10x multiome technology utilized in our study uses nuclei as input and captures nuclear RNA, all previous studies investigating the adult human testis with scRNA-seq technology (Guo, Grow, Yi, et al., 2017; Guo, Grow, Mlcochova, et al., 2018; Guo, Sosa, et al., 2021; Wang, Liu, et al., 2018; Hermann et al., 2018; Xia et al., 2020; Shami et al., 2020; Sohni et al., 2019; Di Persio, Tekath, et al., 2021), except one (Murat et al., 2023), use intact cells as input and capture both nuclear and cytoplasmic RNA. While Murat et al., 2023 have used snRNA-seq to investigate sperm development across multiple species, the authors do not compare their snRNA-seq results to other scRNA-seq studies. The presence or absence of cytoplasmic RNA could, in fact, present a potential challenge in the context of the testis since germ cells in the tissue are connected by cytoplasmic bridges. These are thought to facilitate sharing of cytoplasmic constituents and support synchronous development of the cells (Braun et al., 1989; Morales, Wu, et al., 1998; Huckins, 1971). However, the bridges also lead to mixing of cytoplasmic RNA across different cell types and could affect cell type identification or differential gene expression analyses in scRNA-seq studies.

I first compared the snRNA-seq libraries from our multiome data to scRNA-seq datasets of healthy samples from two published studies - Di Persio, Tekath, et al., 2021 and Guo, Grow, Mlcochova, et al., 2018. I processed all samples in the same manner (for example, the default settings in Cell Ranger v1.2.1 as used by Guo, Grow, Mlcochova, et al., 2018 to map scRNA-seq samples in their analysis discard intronic reads. However, since I include intronic reads in snRNA-seq processing, I re-processed the published datasets from raw data to include these reads and make them comparable to our snRNA-seq datasets; see Methods) and assigned cell types to clusters based on the markers used earlier in Figure 3.6C.

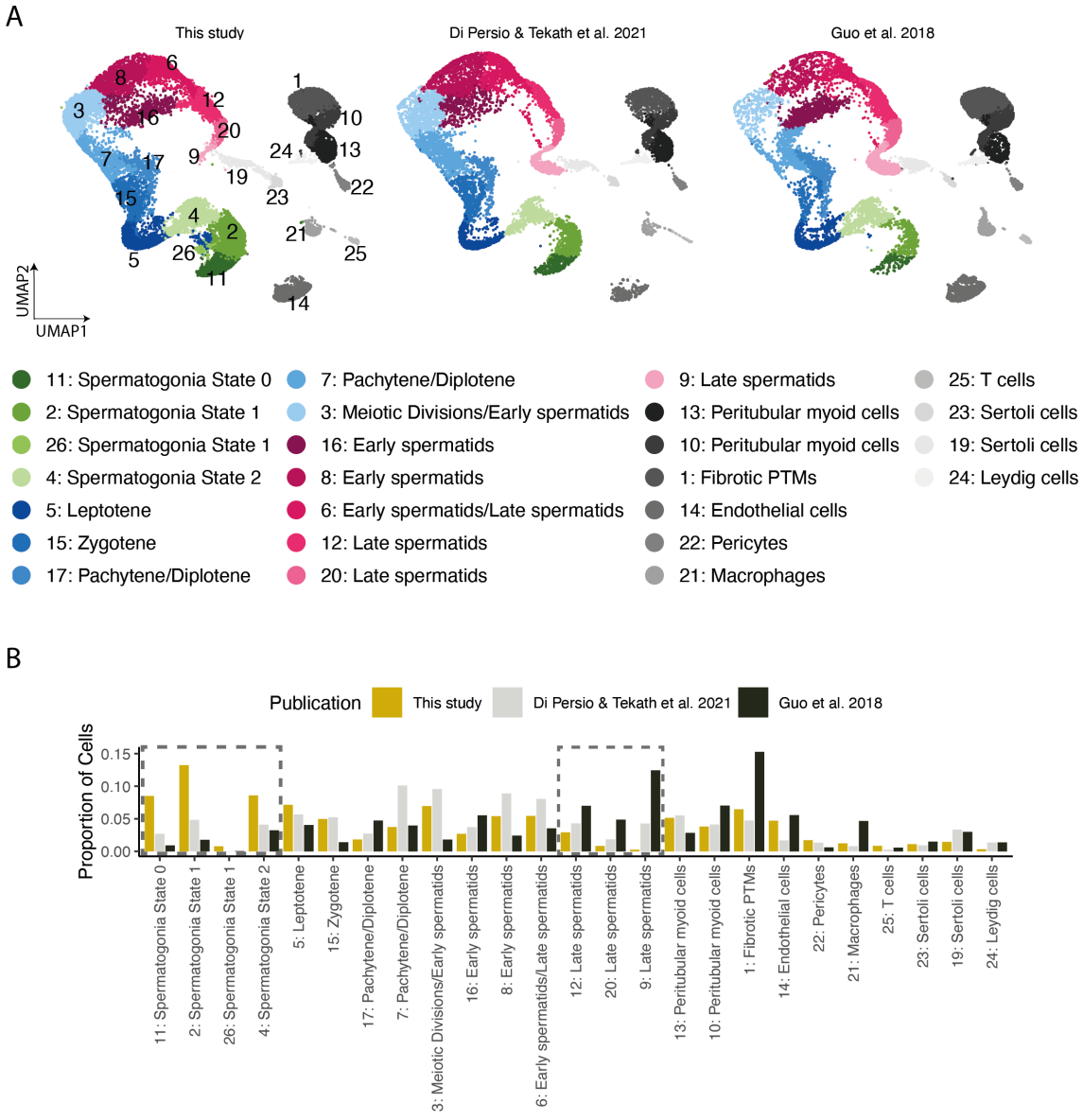



Figure 3.9: ***Comparison of cell proportions between snRNA-seq and scRNA-seq of the adult human testis.*** **(A)** UMAPs of cell clusters obtained from snRNA-seq (This study) and published scRNA-seq from Di Persio, Tekath, et al., 2021 and Guo, Grow, Mlcochova, et al., 2018. **(B)** Proportion of the different cell types captured across the three datasets indicating an enrichment of spermatogonia and a depletion of spermatids in snRNA-seq compared to scRNA-seq datasets. Populations of interest are highlighted within dashed grey boxes.

Overall, I identify all the known cell types across all the studies as seen in the UMAP

(Figure 3.9A). In the comparative analysis, I even identify a few Leydig cells in the snRNA datasets. This is possibly due to the inclusion of more cells of the same type from scRNA-seq datasets, making it easier to obtain a separate Leydig cell cluster during clustering and cell type assignment. Of note, I identify a higher proportion of spermatogonial stem cells in snRNA-seq compared to the published scRNA-seq datasets. Further, in line with our observation of unlysed spermatid heads in the nuclear suspension used to generate the multiome libraries, I capture a lower proportion of haploid spermatids in the snRNA-seq dataset compared to scRNA-seq (Figure 3.9B).

When comparing the number of genes or unique RNA transcripts/unique molecular identifiers (UMIs) captured per study, snRNA-seq captures a higher number of long non-coding RNA (lncRNA) genes and a similar number of protein-coding genes compared to the published scRNA-seq studies (Figure 3.10A, top panel). This is in line with the extensive nuclear functions of lncRNAs (Bridges et al., 2021) and the decreased efficiency of their alternative splicing leading to nuclear retention (Statello et al., 2021). Within the spermatid clusters, scRNA-seq captures a much higher number of UMIs, especially for protein-coding genes (Figure 3.10A, bottom panel). The presence of the RNA-rich chromatoid body in the cytoplasm of differentiating spermatids (Kotaja and Sassone-Corsi, 2007) coupled with the silencing of transcription within the nucleus of these cells could provide a possible explanation for this massive observed difference.

To compare the gene expression patterns between cell types captured in snRNA-seq vs scRNA-seq, I calculated the correlation of average gene expression between the studies for each cell cluster. I found that most clusters in our study show a high significant correlation with the corresponding cluster in both the published scRNA studies (Figure 3.10B). Cluster 26: Spermatogonia State 1 shows the least correlation (< 0.3) with the corresponding cell type in Di Persio, Tekath, et al., 2021. This could possibly be due to the highly limited number of cells assigned to that cluster in the scRNA-seq datasets.

Differential gene expression analysis between the same clusters of snRNA-seq and scRNA-seq datasets reveals limited number of genes that are enriched in either datasets (Figure 3.11). In line with previous observations, late spermatids show a higher number of genes enriched in scRNA-seq than snRNA-seq. Gene Ontology enrichment analysis of these differentially enriched genes result in terms such as "microtubule-based process" for scRNA-seq and "DNA-packaging complex" for snRNA-seq (Figure 3.12). In general, many clusters do not have significant GO terms due to limited number of differentially expressed genes.

Altogether, I observe a high concordance in cell type clusters captured between snRNA-seq and scRNA-seq and their gene expression patterns.

**Contributions** The characterization of the snATAC-seq + snRNA-seq multiome dataset of the adult human testis is a joint venture with Liz Ing-Simmons from the Vaquerizas group. As part of the results included in this chapter, I performed standard quality control, filtering and normalization of the individual replicates and their integration. Liz Ing-Simmons performed clustering and cluster assignment of the integrated dataset, apart from calculating genes enriched in each of the clusters. I performed the com-

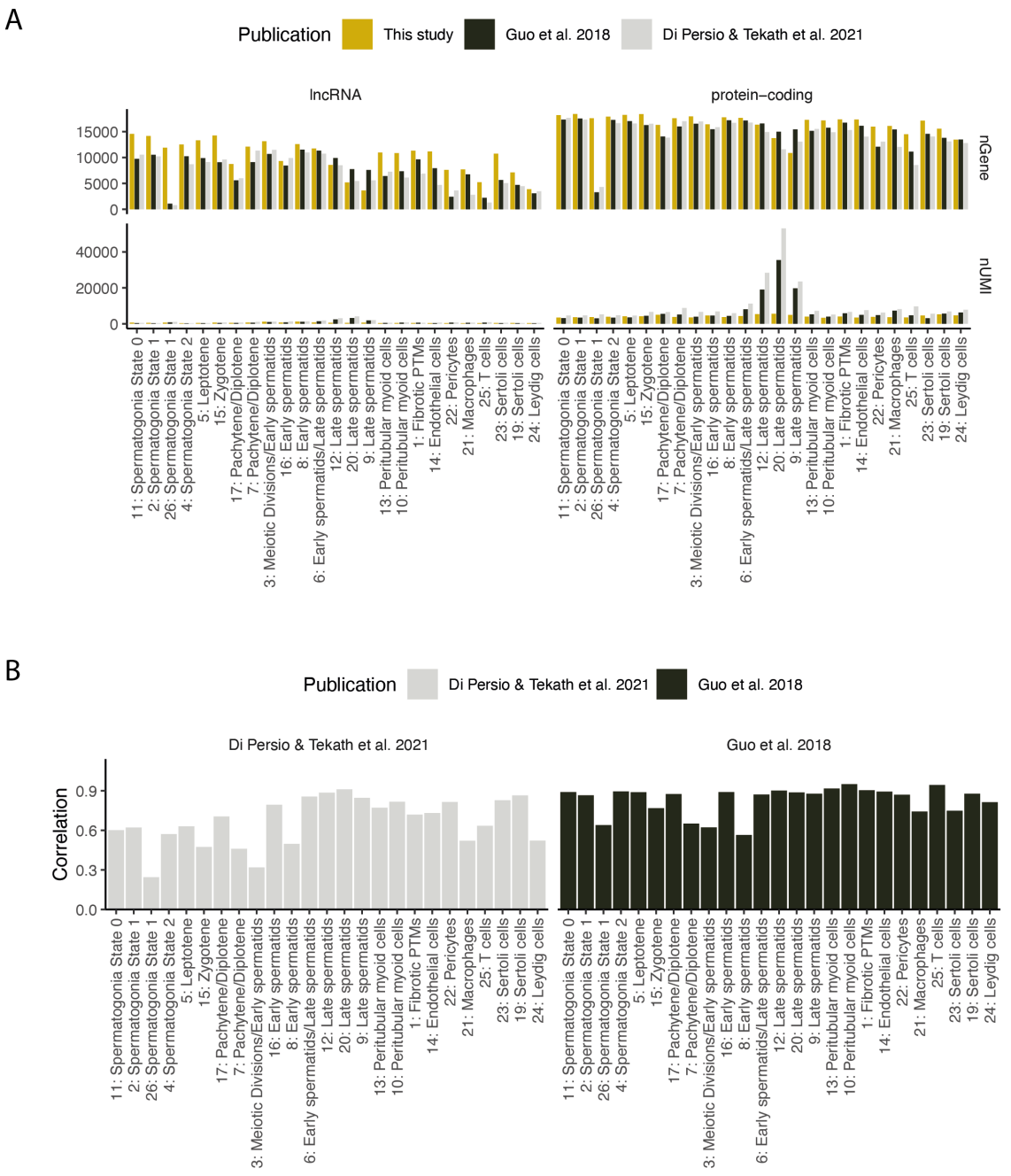



Figure 3.10: ***Comparison of number of genes/UMIs captured between snRNA-seq and scRNA-seq of the adult human testis.*** **(A)** Number of protein-coding or lncRNA genes or unique RNA transcripts (UMIs) captured in each of the datasets in the comparison. **(B)** Correlation of gene expression between This study and Di Persio, Tekath, et al., 2021 or Guo, Grow, Mlcochova, et al., 2018. All values are significant in t-test with Holm–Bonferroni adjustment with p.adj < 0.0001 for Di Persio, Tekath, et al., 2021 and < 0.001 for Guo, Grow, Mlcochova, et al., 2018.

parative study of snRNA-seq vs scRNA-seq. Our collaboration (with subsequent work from Liz Ing-Simmons focussing on the gene regulatory network driving human spermatogenesis) has been submitted for publication titled "Gene regulatory control in human spermatogenesis and infertility" (see Appendix). Specific figures adapted from the manuscript include Figures 3.6, 3.7, 3.8 and 3.9.

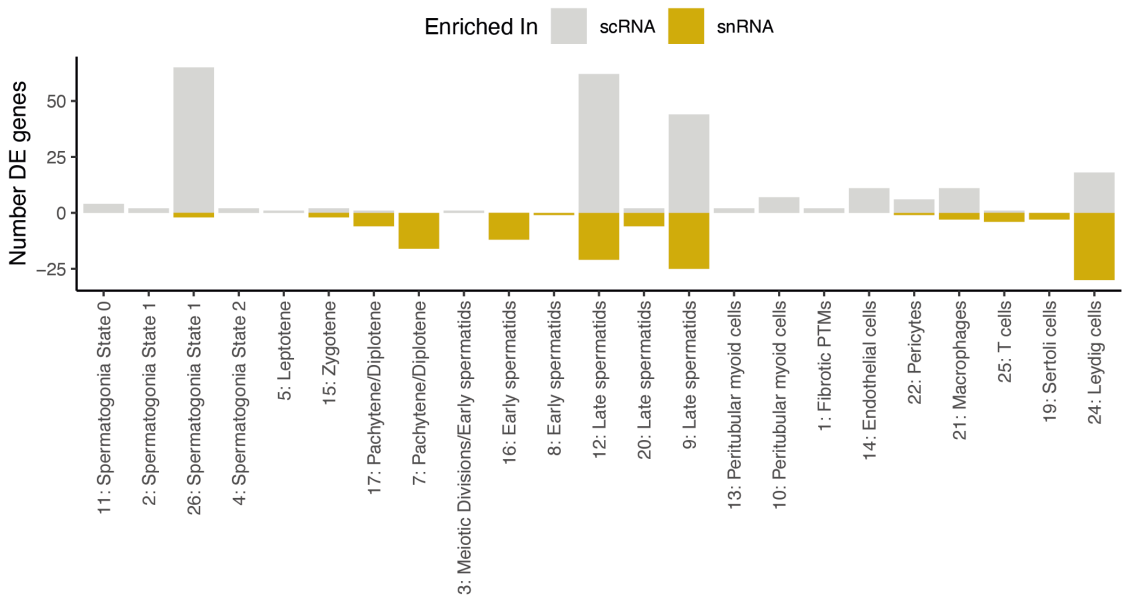



Figure 3.11: ***Number of differentially expressed genes per cluster between snRNA-seq and scRNA-seq of the adult human testis.***

## 3.5 Discussion

In this project, we have produced the first snATAC-seq + snRNA-seq multiome libraries from the same single cells of the adult human testis using the 10x multiome snATAC + Gene Expression platform. Obtaining nuclei from adult human testicular cells to use as input for the platform required extensive optimization of the lysis conditions. CUT&Tag lysis buffer containing 0.01% NP-40 and 0.01% digitonin resulted in maximum number of high quality nuclei compared to *in situ* Hi-C lysis buffer with 0.2% IGEPAL CA-630 or 10x lysis buffer with 0.1% Tween-20, 0.1% NP-40 and 0.01% digitonin. In general, adult human testicular cells are very sensitive compared to commonly used cell lines such as HEK-293T cells or mouse embryonic stem cells (mESCs). While HEK-293T cells and mESCs need to be incubated in lysis buffer for $\sim$13 minutes and 4-5 minutes respectively (data not shown for mESCs), testicular cells start showing signs of over-lysis already after 30 seconds of incubation at 4℃.

The heterogeneity of the testicular cells poses an additional challenge in obtaining a homogenous nuclear suspension from all cell types. While an incubation period of 30 seconds at 4℃ isolates nuclei from most cells, a few unlysed elongating spermatid heads still remain. In an attempt to obtain a more homogenous nucleus solution, I fixed the cells with MeOH prior to nuclei isolation due to successful and promising results for scRNA-seq (CG000136, Revision E) and snATAC-seq (email correspondence with 10x technical support team) library generation from MeOH-fixed cells respectively. While I could not lyse MeOH-fixed cells to release their nuclei, I proceeded with multiome library generation with MeOH-fixed cells itself as input since MeOH fixation additionally permeabilizes the cells and could thus allow entry of transposition reagents into the nucleus of the cell. Unfortunately, the snATAC-seq library generated thus did not show the expected nucleosomal patterning for successful transposition in

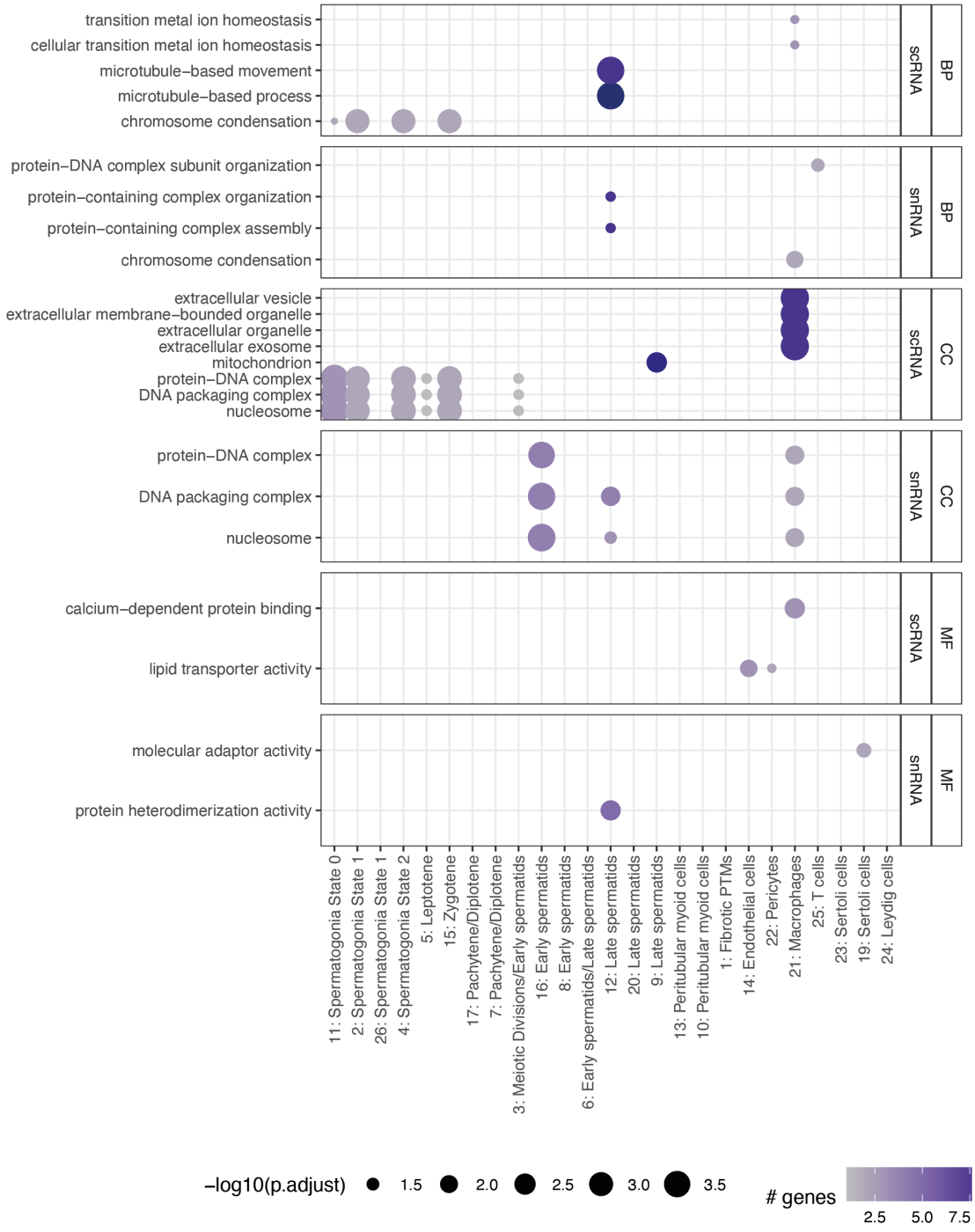



Figure 3.12: ***Gene Ontology enrichment analysis of differentially expressed genes between snRNA-seq and scRNA-seq of the adult human testis.***

their fragment size distribution. Furthermore, when testing nuclei isolation from cryopreserved, snap-frozen and fresh mESCs, fresh cells produced the maximum number of high quality nuclei (data not shown). Hence, I proceeded to generate the multiome libraries using nuclei isolated from fresh adult human testicular cells. Since we capture fewer elongating spermatids in our nuclear suspension, the observed proportions of cell types in our dataset do not reflect the true composition of the tissue.

To address the challenge of heterogeneity in lysis conditions, prior to nuclei isolation, one could sort the testicular cells based on DNA content to separate the elongating spermatids from rest of the population and treat them to stronger lysis conditions. One could also try other fixation agents like dithio-bis(succinimidyl propionate) (DSP or Lomant's Reagent) (Murat et al., 2023) or the more commonly used Paraformaldehyde (PFA). While 10x Genomics did not have a successful working protocol for scRNA-seq library generation from PFA-fixed cells at the time of optimization, they have since released the "Single Cell Fixed RNA Profiling (FRP) workflow" to achieve the same. With respect to the impact of PFA-fixation on snATAC-seq libraries, Lareau et al., 2021 have generated successful mtscATAC-seq, which simultaneously captures the mitochondrial genome and high-quality chromatin accessibility data from PFA-fixed cells. Furthermore, the 10x technical support team have also observed promising results in internal pilot experiments of snATAC-seq library generation with nuclei isolated from PBMCs fixed with 0.1% - 0.5% PFA (personal correspondence). Thus, additional optimization of the protocol may be beneficial in capturing more spermatids.

The necessity to use fresh samples in our protocol and the inherent fragility of the cells make it challenging to expand this technique to a vast majority of samples present in long-term storage of fixed/cryopreserved/snap-frozen conditions. Furthermore, in Murat et al., 2023's comparative study of testis transcriptomes across multiple species using snRNA-seq, the authors needed to fix therian (eutherian and marsupial) samples, but not platypus and chicken, with DSP prior to nuclei isolation. Hence, additional optimization of the protocol might also be required based on the origin species of the tissue.

Our multiome libraries capture almost all known cell types from the adult human testis with high similarity between the biological replicates. However, we cannot identify a unique cluster of Leydig cells in our dataset, possibility due to the enrichment of intratubular cells over those present in the interstitial compartment during the preparation of single cell suspension. The Leydig cells were found to be distributed amongst other somatic cell clusters. As with published studies (Guo, Grow, Mlcochova, et al., 2018; Wang, Liu, et al., 2018; Hermann et al., 2018; Sohni et al., 2019; Shami et al., 2020), we also see stage-specific gene expression patterns that include both known and potentially new marker genes. Additionally, our matched snATAC-seq data captures distal non-promoter peaks that show stage-specific accessibility and could be putative enhancers or testis-specific promoters. Importantly, the joint analysis helped us resolve the identity of the meiotic division cluster, which occurs at end of Meiosis I. The cluster presented overall low expression of marker genes in snRNA-seq but showed a clear bimodal distribution of reads associated with chrX in the snATAC-seq data representing the distribution of sex chromosomes into distinct cells at this stage.

In comparison to published scRNA-seq datasets, our study captures more spermatogonia and fewer late spermatid populations. The higher proportion of spermatogonia can aid in dissecting the dynamics of spermatogonia self-renewal and differentiation (discussed in Chapter: Spermatogonia Dynamics). The lower proportion of spermatid clusters is in line with the presence of unlysed elongating spermatid heads in the nuclear suspension used to generate the multiome libraries and could limit our analyses of spermiogenesis. For example, we capture sufficient cells and UMIs associated with spermatid clusters to investigate gene expression differences between

X- and Y-spermatids (see Chapter: Comparison of X- and Y-spermatids). However, coupled with the inherent sparsity of snATAC-seq data, we are unable to get adequate spermatid-specific snATAC-seq reads to identify the precise genomic location of nucleosomes that could help us study the dynamics of histone-to-protamine exchange (data not shown).

Surprisingly, snRNA-seq captures a higher number of genes than scRNA-seq. In contrast, previous work from 10x Genomics detected 20-40% fewer genes in snRNA-seq datasets of human PBMCs compared to scRNA-seq datasets of the same cells (10xGenomics, n.d.). This nuclear enrichment is especially strong for lncRNAs which could be due their lesser efficiency of alternative splicing (Statello et al., 2021) or their targeted functions within the nucleus (Bridges et al., 2021). scRNA-seq, on the other hand, captures a much higher number of UMIs associated with the late spermatid clusters. Since the cells are mostly transcriptionally silent at this stage, it is expected that there would be fewer associated UMIs in snRNA-seq. The much higher number in scRNA-seq could also partly be a technical consequence of difference in cell numbers captured, wherein a higher number of spermatids in scRNA-seq datasets also leads to a higher number of UMIs associated with it. This is also observed in cluster 26: Spermatogonia State 1, where fewer cells in scRNA-seq is associated with fewer detected genes (Figure 3.10) and higher number of differentially enriched genes relative to snRNA-seq (Figure 3.11). However, to account for such misleading observations, the number of UMIs have already been normalized for both sequencing depth and cell numbers. Alternatively, spermatids are known to store transcripts from previous stages for translation in late spermiogenesis (Geremia et al., 1977). This could also explain the occurrence of GO terms like "microtubule-based process" or "mitochondrion" in spermatid genes enriched in scRNA-seq (Figure 3.12), as these organelles undergo extensive changes during spermiogenesis. Lastly, the scRNA-seq transcripts could be associated with the cytoplasmic RNA-rich chromatoid body found in post-meiotic germ cells. In that case, the higher number of UMIs associated with the last spermatid cluster (Cluster 9: Late spermatids) could indicate a dearth of late stage spermatids in the scRNA-seq datasets as chromatoid bodies are found to disappear in mature elongated spermatids (Kotaja and Sassone-Corsi, 2007).

A key limitation of any transcriptomic study, including this, is the lack of definitive functional inference without validation at the protein level. Studying the single cell suspension of the testis further precludes an understanding of its spatial organization which could be captured by the newly-developed spatial transcriptomics technologies (Chen, Murray, et al., 2021; Vandereyken et al., 2023). Furthermore, post-transcriptional mechanisms regulating spermatogenesis such as uridylation (Morgan, Kabayama, et al., 2019) and $m^6A$ modification (S. Liu et al., 2021; Steger, 1999; Morgan, Kumar, et al., 2021), will not be captured in our study. Despite these limitations, we can use this data to infer the gene regulatory networks driving spermatogenesis (see Appendix). In summary, we provide a comprehensive single nucleus map of chromatin accessibility and gene expression during healthy human spermatogenesis, which can be further be used to inform the interpretation of genes and genomic variation implicated in male infertility.

# Chapter 4

# Dynamics of spermatogonia homeostasis in the adult human testis

## 4.1 Introduction

The spermatogenesis process is sustained by a population of spermatogonial stem cells that reside in the periphery of the seminiferous tubule. In humans, histological analysis has revealed two types of undifferentiated spermatogonia (Types $A_{dark}$ and $A_{pale}$), and one type of differentiating spermatogonia (Type B). $A_{dark}$ is believed to be the reserve and $A_{pale}$ the active stem cell population in the adult human testis (Clermont, 1969; Clermont, 1972; Van Alphen et al., 1988; Caldeira-Brant et al., 2020).

However, recent advances in single cell technology has facilitated the discovery of new subgroups based on their transcriptomic profiles, including undifferentiated States 0-1 and differentiating States 2-3 (Guo, Grow, Mlcochova, et al., 2018; Guo, Nie, et al., 2020; Guo, Sosa, et al., 2021; Di Persio, Tekath, et al., 2021; Shami et al., 2020; Sohni et al., 2019). State 0 has been identified as the most naive stem cell population in the testis (Sohni et al., 2019; Guo, Grow, Mlcochova, et al., 2018; Shami et al., 2020; Tan et al., 2020) with other studies discovering two more subpopulations of State 0, termed State 0A and State 0B (Sohni et al., 2019; Di Persio, Tekath, et al., 2021). However, there is no direct correspondence between the states identified in microscopic studies and scRNA-seq studies (Jan et al., 2017; Di Persio, Tekath, et al., 2021). Spatial transcriptomics also captured distinct spatial clusters within the tubule that are occupied by the different subpopulations, and novel signalling interactions that regulate the stem cell niche (Rajachandran et al., 2023).

Despite extensive transcriptomic characterization of the spermatogonial compartment, very little is known about the gene regulatory mechanisms maintaining stem cell homeostasis. Bulk ATAC-seq studies of FACS-sorted spermatogonia enriched for stage-specific markers (undifferentiated marker SSEA4 and differentiating marker KIT) have identified the enrichment of DMRT1, NFYA/B, CTCFL motifs in undifferentiated spermatogonia (Guo, Grow, Yi, et al., 2017). However, they do not find

differences in chromatin accessibility profiles between the undifferentiated and differentiating states (Guo, Grow, Mlcochova, et al., 2018). snATAC-seq of the human testis has also captured motifs enriched in certain stages across germ cell development. However, the analysis fails to discern the spermatogonial compartment at the resolution of the scRNA-seq studies described earlier (Wu et al., 2022).

In this chapter, I present a detailed description of the gene regulatory landscape across the human spermatogonial compartment at high resolution. Apart from identifying the known stages of States of 0-3, I also describe eight new subclusters identified by joint analysis of snRNA-seq and snATAC-seq. This includes a subpopulation at the transition of undifferentiated to differentiating spermatogonia, whose gene expression correlates with the former state and motif accessibility correlates with the latter. I also show that State 0B and State 0A lie on two diverging branches of UMAP connecting State 0 to State 1, where State 0B has a higher proportion of proliferative cells and a higher percentage of mitochondrial transcripts compared to State 0A. Furthermore, both subgroups exhibit different niche communications including interactions that are misregulated in infertility.

## 4.2 snATAC-seq + snRNA-seq multiome identifies new subclusters of spermatogonia in the adult human testis

The snATAC-seq + snRNA-seq multiome dataset of the adult human testis generated in this study captures a higher proportion of spermatogonia compared to other published scRNA-seq datasets (Chapter: Generation and characterization of joint snATAC-seq + snRNA-seq multiome datasets of the adult human testis; Figure 3.9B). Thus, the multiome data provides an avenue for detailed characterization of the stem cell compartment of the testis. The WNN UMAP of the whole testis captures two diverging branches of undifferentiated spermatogonia (State 0A and State 0B) that merge together before forming the differentiating spermatogonia clusters and subsequently following a linear trajectory of development through meiosis and spermatid differentiation (Figure 4.1A). To dissect the dynamics of spermatogonia development in more detail, I subset the spermatogonia clusters from the whole testis dataset and re-clustered them using the weighted nearest-neighbour (WNN) approach. This gave me 13 clusters in total (Figure 4.1B) and I assigned the cell types based on the known markers given in Figure 4.2A (Di Persio and Neuhaus, 2023). *DDX4* is a pan-germ cell marker while *MAGEA4* is a pan-spermatogonia marker in humans (Landry et al., 2000). *PIWIL4* is an argonaute protein that was found enriched in State 0 spermatogonia, the most naive population identified in the human testis (Guo, Grow, Mlcochova, et al., 2018). Later studies (Di Persio, Tekath, et al., 2021; Sohni et al., 2019) identified two more subpopulations of State 0, termed State 0A and State 0B, enriched for *FGFR3* and *NANOS2* respectively. *GFRA1* is a marker for State 1 undifferentiated spermatogonia. Within the differentiating spermatogonia clusters, *DMRT1, DNMT1, KIT, MKI67* and *SOHLH2* are markers for State 2 spermatogonia while *REC8* and *STRA8* are markers for State 3 spermatogonia.

Based on marker gene expression, I identify seven clusters of undifferentiated sper-

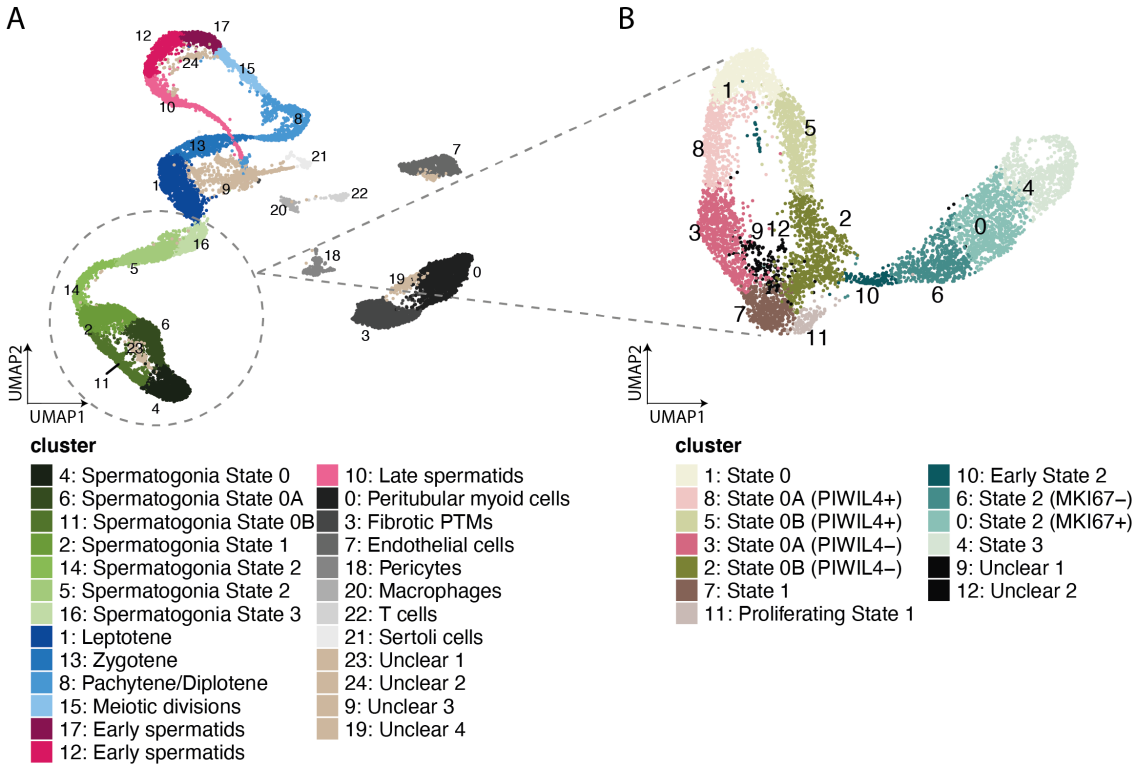



Figure 4.1: ***snATAC-seq + snRNA-seq multiomic characterization of spermatogonial compartment in the adult human testis.*** WNN UMAP representation of single-nucleus multiomics data from **(A)** adult human testis and **(B)** re-clustering of the spermatogonia clusters that were subset from the whole testis dataset. "Unclear" clusters that could not be identified based on marker gene expression and were excluded from further analysis.

matogonia (clusters 1, 8, 5, 3, 2, 7, 11) and four clusters of differentiating spermatogonia (clusters 10, 6, 0, 4). I also identify two clusters that could not be assigned based on known markers and hence were discarded from further analyses (cluster 9 and 12) (Figure 4.1B and 4.2A).

Apart from the known subtypes of spermatogonia, I identify eight new clusters that were previously uncharacterized. Within the State 0A cluster, I identify two populations, one enriched for and one lacking the expression of *PIWIL4*. These were termed 8: State 0A (PIWIL4+) and 3: State 0A (PIWIL4-) respectively. State 0B also separates into PIWIL4+ (5: State 0B (PIWIL4+)) and PIWIL4- (2: State 0B (PIWIL4-)) clusters (Figure 4.2B, left panel). Within State 1 spermatogonia, I identify a subgroup of proliferating cells as identified by the expression of the proliferation marker *MKI67* (11: Proliferating State 1) (Figure 4.2B, middle panel). Lastly, State 2 cells separate into three clusters including 10: Early State 2, 6: State 2 (MKI67-) and 0: State 2 (MKI67+). 10: Early State 2 only shows an enrichment of *DMRT1* and not *KIT*. Clusters 6: State 2 (MKI67-) exhibits lesser expression of MKI67 than cluster 0: State 2 (MKI67+) (Figure 4.2B, right panel). Thus classified, snATAC-seq + snRNA-seq multiome analysis identifies new subgroups of spermatogonia in the adult human testis.

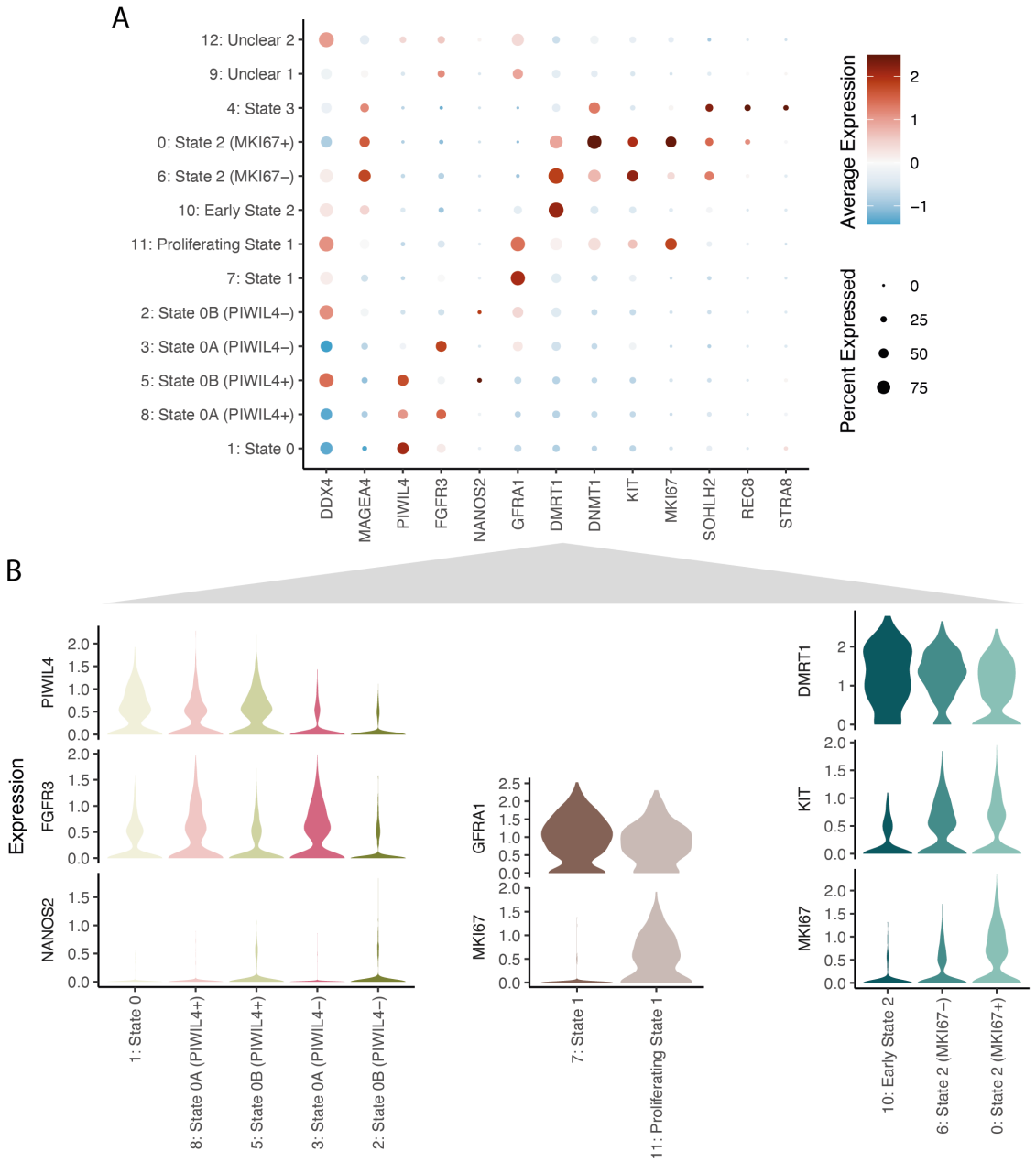



Figure 4.2: ***Marker gene expression across the spermatogonia subclusters in the adult human testis.*** **(A)** Marker gene expression across clusters used to assign cluster identity. Colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression. **(B)** Violin plots of marker gene expression used to identify the new subclusters.

## 4.3 Gene expression profile of Early State 2 resembles undifferentiated spermatogonia

Upon identifying new subtypes of spermatogonia in the adult human testis, I first characterized their gene expression patterns. A heatmap of the top 100 genes enriched per cluster shows cell type-specific gene expression patterns (Figure 4.3A).

The top three genes enriched in each of the new spermatogonia subtypes have been annotated in the heatmap in the respective colours of the clusters. Many of these

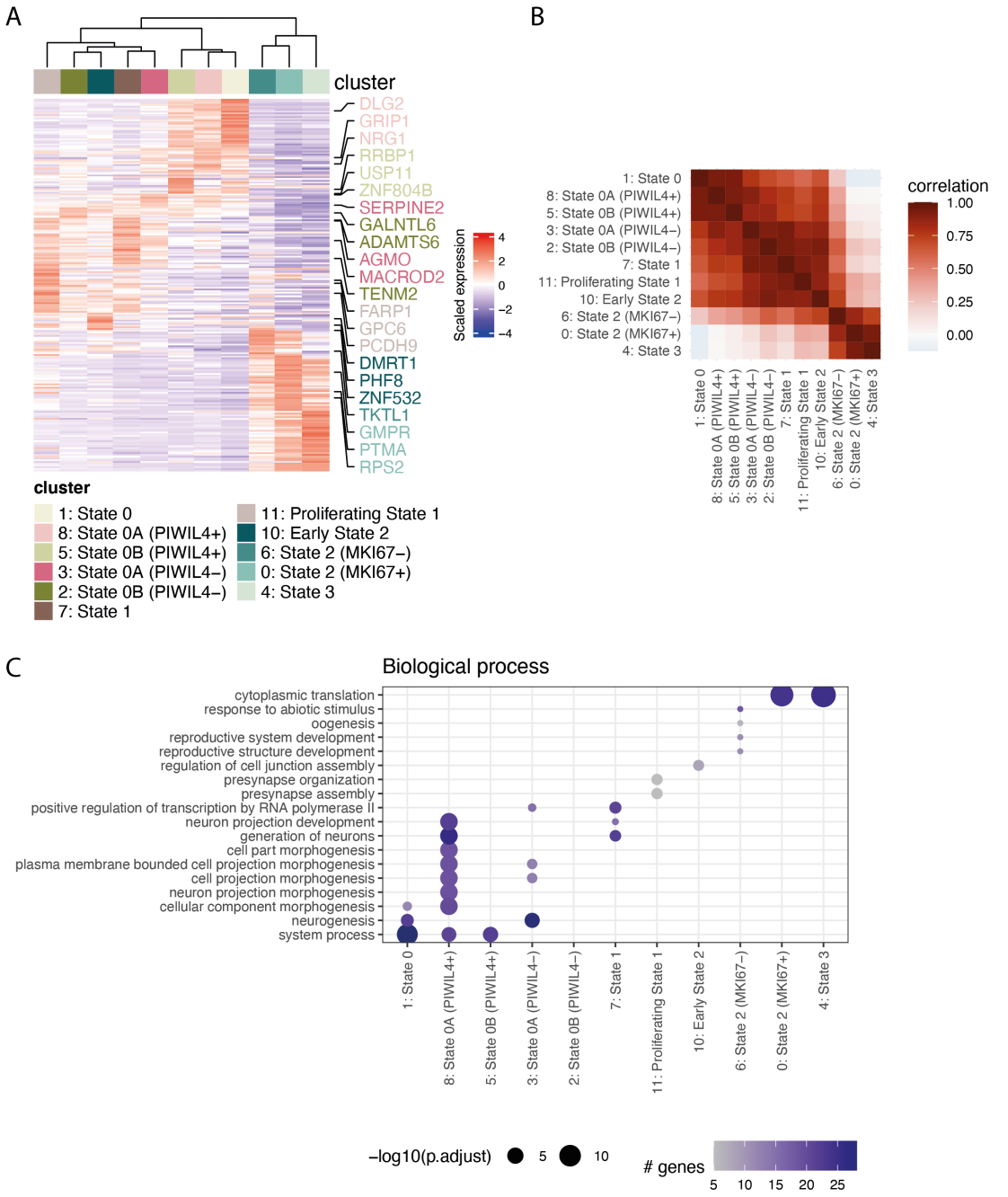



Figure 4.3: ***Gene expression characterization of the spermatogonia subclusters in the adult human testis.*** **(A)** Hierarchical clustering of gene expression heatmap across the spermatogonia subclusters. Genes are ordered by the cluster in which the gene has max expression. The top three genes enriched for new subclusters identified are labelled on the right-hand side. The gene labels are coloured by the cluster of enrichment. **(B)** Heatmap of Spearman correlation coefficient of top 100 genes enriched per cluster. All correlation coefficients are significant with p value < 0.05. **(C)** Gene Ontology (Biological Process) terms that are significantly enriched in marker genes for each cluster. The top most significant term per cluster was selected for inclusion in the plot. Size represents -log10 of the BH-adjusted p-value for enrichment of the term in the given gene set, and colour represents the number of DE genes associated with that term. Absent dots indicate no significant enrichment of that term in the given gene set.

genes are found associated with male subfertility. *NRG1* is enriched in 8: State 0A (PIWIL4+) and belongs to the epidermal growth factor family. *NRG1* promotes

clonal development of rat differentiating spermatogenic cells *in vitro* (Chapman et al., 2015), and induces spermatogonial proliferation and meiotic initiation in organ cultures mouse testis (Zhang, Eto, et al., 2011). Furthermore, *NRG1* mutant mice show dramatically reduced Leydig cell proliferation in the infant testes (Umehara, Kawashima, et al., 2016). *GALNTL6* is a glycosyltransferase-like gene and is enriched in 2: State 0B (PIWIL4-). It has been associated with inbreeding depression of sperm concentration in bulls (Ghoreishifar et al., 2023). Furthermore, mutation in a related gene *GALNTL5* results in reduced sperm motility and male infertility (Takasaki et al., 2014). SNPs in *PCDH9*, a protein of the cadherin superfamily enriched in 11: Proliferating State 1 and in *MACROD2*, a deacetylase enriched in 3: State 0A (PIWIL4-), have been associated epididymis dysfunction and male infertility (Bischof et al., 2013). *TKTL1*, a transketolase gene enriched in 6: State 2 (MKI67-), has been proposed to be used as a biomarker for the state of spermatogenesis in patients with NOA, since the protein was consistently detected in normal seminal plasmas, but not in NOA, OA and vasectomized men (Rolland et al., 2013). It is also found to be significantly downregulated in low-fertility bulls (Talluri et al., 2022). Lastly, the ribosomal protein RPS2 enriched in 6: State 2 (MKI67-) is associated with altered protein-protein interactions in NOA (Aras-Tosun, 2023).

Hierarchical clustering of the gene expression heatmap reveals two major groups - one mostly consisting of the undifferentiated spermatogonia clusters and the other mostly consisting of the differentiating spermatogonia clusters. Within the undifferentiated spermatogonia group, PIWIL4+ cell types (1: State 0, 8: State 0A (PIWIL4+) and 5: State 0B (PIWIL4+)) form a subgroup. Interestingly, 10: Early State 2, a differentiating spermatogonia cluster as identified by marker gene expression (Figure 4.2A), clusters with the undifferentiated spermatogonia cell types. These observations are also supported by the correlation coefficient values for pairs of spermatogonia subclusters computed on the top 100 cluster-specific genes (Figure 4.3B).

Gene Ontology enrichment analysis results in GO terms such as "cytoplasmic translation" associated with the clusters 0: State 2 (MKI67+) and 4: State 3, which possibly captures the increase in cell activity in preparation for the mitosis-to-meiosis transition (Figure 4.3C).

Thus, differential gene expression analysis of the adult human spermatogonial compartment identifies cell type-specific expression patterns, including genes associated with subfertility. We also capture a subgroup of Early State 2 spermatogonia, which is considered to be a differentiating state based on known marker expression but presents a gene expression profile that correlates more with the undifferentiated spermatogonia cell states.

## 4.4 Transcription factor motif activity profile of Early State 2 resembles differentiating spermatogonia

The snATAC-seq libraries of the multiome dataset enables us to examine the chromatin accessibility landscape that contributes to the cell type-specific gene expression patterns observed within the spermatogonial compartment. The dynamic changes in the cluster-specific number of open or closed peaks across the genome reveals a

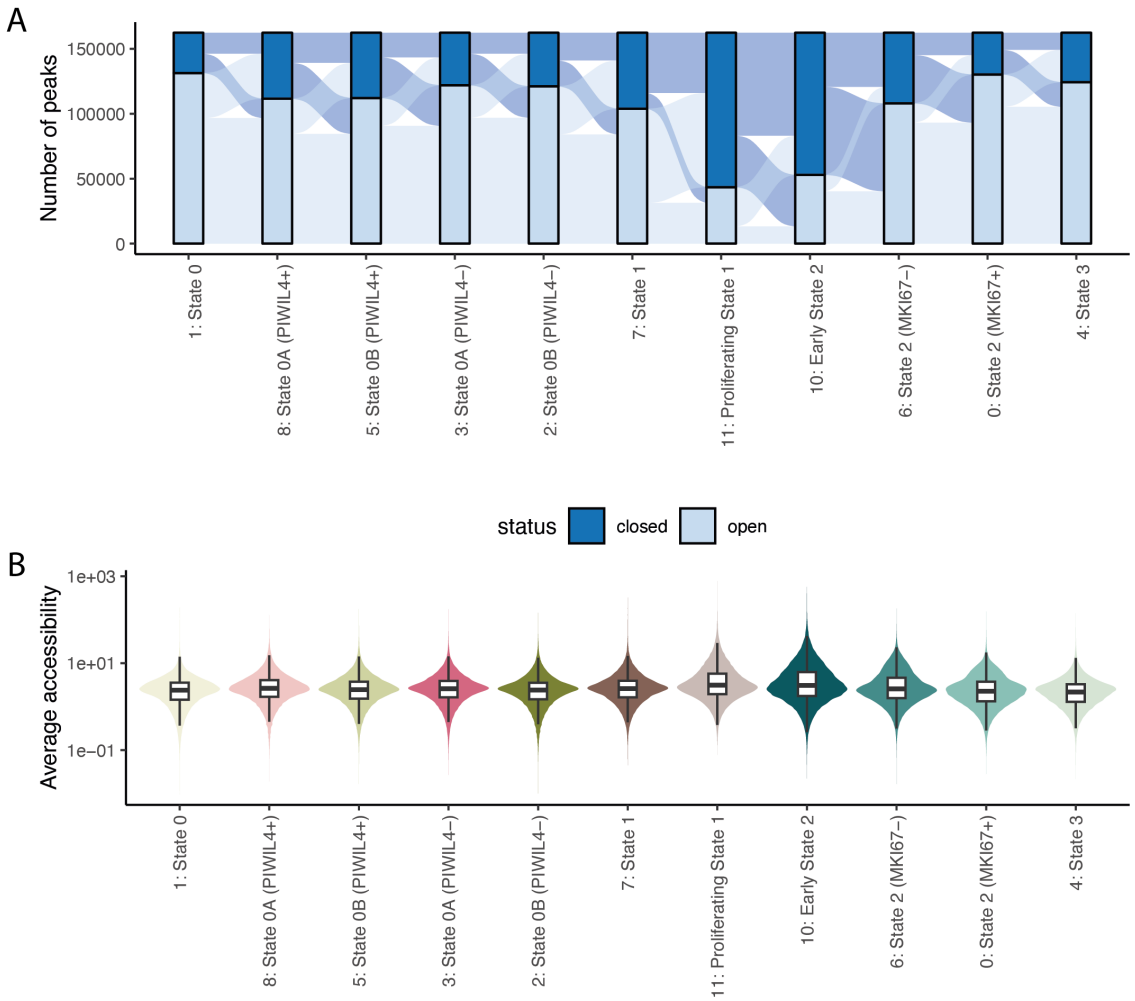



Figure 4.4: ***Genome-wide differential chromatin accessibility characterization of the spermatogonia subclusters in the adult human testis.*** **(A)** Alluvial plot of the number of open or closed peaks for each cluster. A peak was classified as "open" if the average accessibility of the peak across all cells in the cluster was higher than 0. **(B)** Distribution of average accessibility of peaks in the cluster.

pattern of genome closing and re-opening as spermatogonia transition from an undifferentiated to a differentiating state (Figure 4.4A). Clusters 11: Proliferating State 1 and 10: Early State 2 present the lowest number of open peaks. However, genome-wide average accessibility across all the spermatogonia clusters is comparable, with clusters 11: Proliferating State 1 and 10: Early State 2 even showing slightly higher accessibility than the others (Figure 4.4B). Taken together, clusters at the juncture between undifferentiated and differentiating spermatogonia appear to have a very selective number of open peaks, but these peaks exhibit overall high accessibility. Alternatively, they could represent a transitioning state where the predominant number of peaks are either decreasing or increasing in accessibility, with neither group of peaks having reached their minimum or maximum accessibility compared to the flanking clusters.

Focussing on chromatin accessibility within known motifs of TFs, termed motif activity, detects stage-specific enrichment patterns. The top five motifs enriched in each of the new spermatogonia subtypes have been annotated in the heatmap in the respective

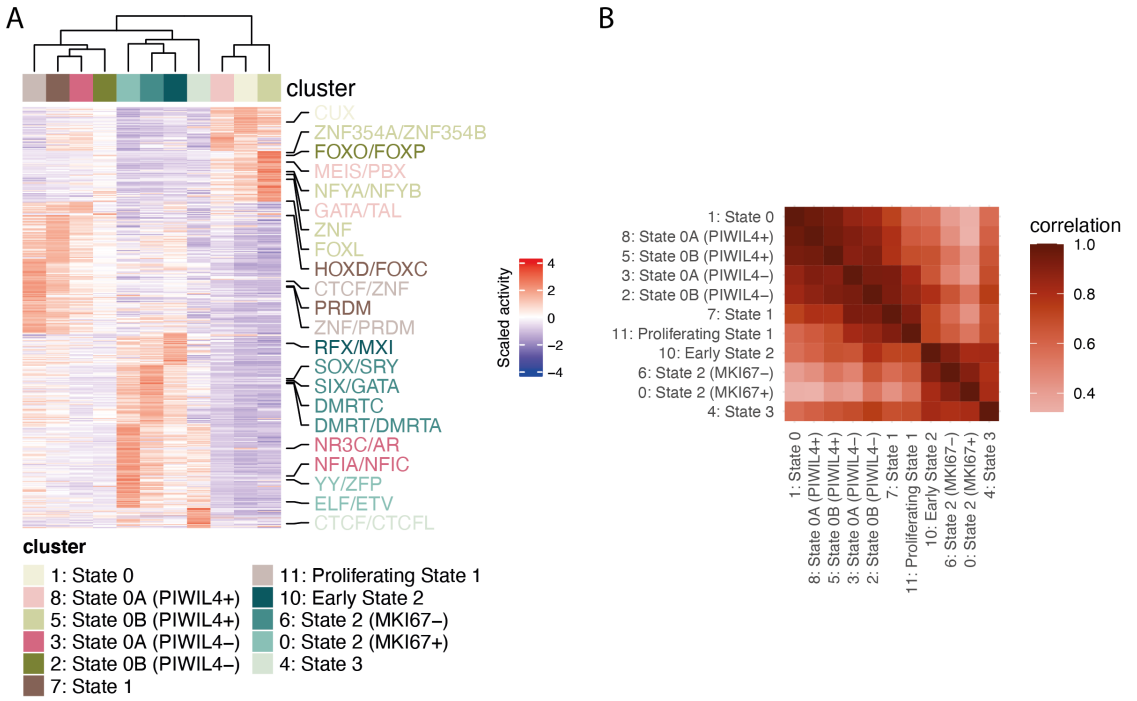



Figure 4.5: ***Differential motif activity characterization of the spermatogonia subclusters in the adult human testis.*** **(A)** Hierarchical clustering of heatmap of motif activity across the spermatogonia subclusters. Motifs are ordered by the cluster in which the motif has maximum activity. The top four motifs enriched for different cell types are labelled on the right-hand side. The motif labels are coloured by the cluster of enrichment. **(B)** Heatmap of correlation coefficient of top 100 motifs enriched per cluster. All correlation coefficients are significant with p adjusted value < 0.01.

colours of the clusters (Figure 4.5A). Interestingly, 1: State 0 shows high activity of the CUX motif. Inactivating mutations in CUX1, a TF that binds to this motif, has been implicated in clonal expansion and neoplasms in the hematopoietic system (An et al., 2018). Hence, CUX motif activity in the human spermatogonial compartment is possibly associated with maintaining 1: State 0 in a quiescent state. Also notably, 3: State 0A (PIWIL4-) shows enriched motif activity for the NR3C/AR motif, which is bound by the androgen receptor, a known regulator of sperm and testicular development (R.-S. Wang et al., 2009). Lastly, as expected, the motifs DMRTC and DMRT/DMRTA possibly bound by DMRT1, a marker for differentiating spermatogonia, show high activity in the differentiating spermatogonia cluster 6: State 2 (MKI67-). We also identify motifs such as NFYA/NFYB and CTCF/CTCFL that have been reported to show enriched activity in other cell types in the testis (Wu et al., 2022).

Hierarchical clustering of the top 100 accessible motifs per cluster reveals 3 main groups (Figure 4.5A). Similar to the gene expression profiles seen in Figure 4.3A, PIWIL4+ cell types (1: State 0, 8: State 0A (PIWIL4+) and 5: State 0B (PIWIL4+)) form one group. Here, PIWIL4- (3: State 0A (PIWIL4-) and 2: State 0B (PIWIL4-)) and State 1 clusters (7: State 1 and 11: Proliferating State 1) form another group. Surprisingly, 10: Early State 2 now clusters with the differentiating spermatogonia cell types. This is also evident in the correlation coefficient values for pairs of spermatogonia subclusters computed on the top 100 motifs with cluster-specific activity (Figure 4.5B). Furthermore, Figure 4.5B reveals a high correlation of 2: State 0B (PIWIL4-) with all the other cell types. However, Figure 4.5A presents no specific enrichment or

depletion of motif accessibility in 2: State 0B (PIWIL4-). This lack of preferential motif accessibility within the cluster probably drives the observed correlation with the other clusters.

Overall, PIWIL4+ clusters 1: State 0, 8: State 0A (PIWIL4+) and 5: State 0B (PIWIL4+) have more motifs with higher accessibility compared to the other clusters (Figure 4.6A). However, even the top accessible motifs in each of these clusters, CUX:CUT in 1: State 0, MEIS/PBX in 8: State 0A (PIWIL4+) and NYFA/NFYB:CBF/NFY in 5: State 0B (PIWIL4+) (Figure 4.6B) show only a slightly stronger footprint in the respective cell types of enrichment compared to the others.

Hence, while the spermatogonia clusters show cell type-specific motif accessibility, this does not translate to an equally strong cell type-specific transcription factor binding.

## 4.5 State 0B exhibits a higher enrichment of mitochondrial transcripts and proliferative cells compared to State 0A

After characterizing the global gene expression patterns and chromatin accessibility landscape of the entire spermatogonial compartment, I explored the possible causes underlying the unique arrangement of State 0A and State 0B clusters on either side of the donut-shaped organization of the undifferentiated spermatogonia in the WNN UMAP. I first inspected if confounding effects such as cell stress or cell cycle could lead to such an arrangement. A high percentage of mitochondrial reads could signify metabolically active or low quality/apoptotic cells. Another aspect indicative of low quality cells is the fraction of reads in peaks (FRiP) associated with snATAC-seq data. Since FRiP is a measure of usable reads in significantly enriched peaks, a lower value of FRiP implies a poorer quality of ATAC-seq data.

Figure 4.7A reveals that 5: State 0B (PIWIL4+) and 2: State 0B (PIWIL4-) have a significantly higher percentage of mitochondrial transcripts than 8: State 0A (PIWIL4+) and 3: State 0A (PIWIL4-) respectively. This possibly indicates more stressed/low quality cells in State 0B. However, the State 0B clusters simultaneously have a significantly higher FRiP value compared to the respective State 0A clusters (Figure 4.7B). Thus, neither State 0A nor State 0B consistently show properties of low quality cells compared to the other.

Next, I inspected if possible differences in cell cycle phase could drive the divergence of State 0A and State 0B into separate branches of the WNN UMAP. For this, I used Tricycle which estimates cell cycle position of a single cell along a circular and continuous scale from 0 to $2\pi$. Approximately, $0.5\pi$ corresponds to the start of the S stage, $\pi$ to be the start of G2M stage, $1.5\pi$ to be the middle of M stage, and $1.75\pi$-$0.25\pi$ to be G1/G0 stage (Zheng et al., 2022).

The undifferentiated spermatogonia clusters of State 0 and State 1 (except 11: Proliferating State 1) are enriched in cells at G1/G0 phase (Figure 4.8A and B). Tricycle analysis neatly captures cells of 11: Proliferating State 1 in the S phase. While 10: Early State 2 is mostly enriched for cells at G1/G0, the subsequent differentiating

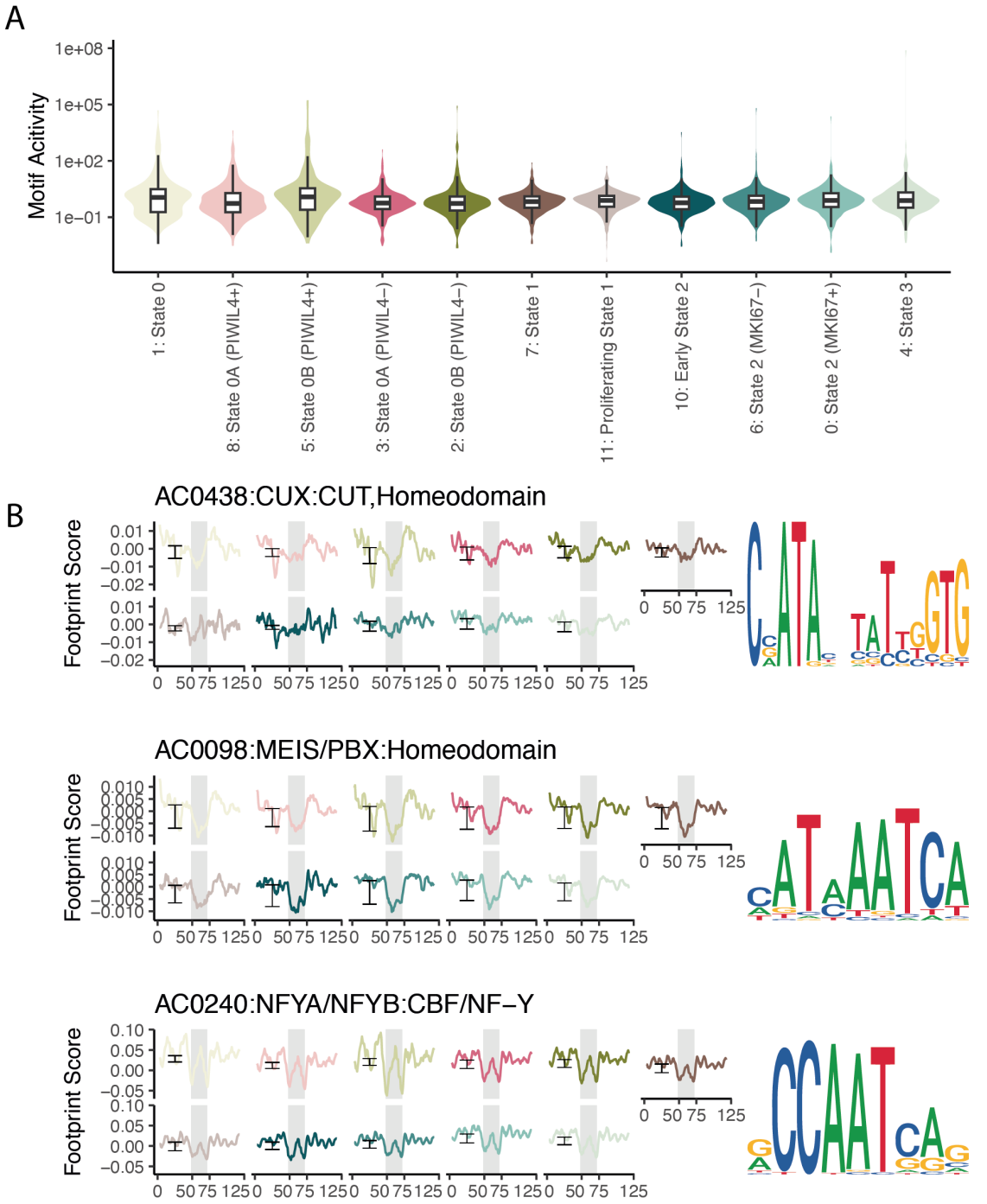



Figure 4.6: ***Transcription factor footprinting profiles across the spermatogonia subclusters in the adult human testis.*** **(A)** Distribution of average activity of motifs in the cluster. **(B)** The TOBIAS footprint profiles for all spermatogonia subclusters (left) and position weight matrix (PWM) (right) for motifs with the highest activity in 1: State 0 (top), 8: State 0A (PIWIL4+) (middle) and 5: State 0B (PWIIL4+) (bottom). Grey-shaded area represents the centre of the motif and the error bars depict the footprint depth or the difference in accessibility between centre and flank of the motif binding region. The profiles of these motifs indicate similar TF binding between the cluster with the highest motif activity and the other subclusters.

spermatogonia clusters are shown to be enriched in cells in the proliferative phases of S and G2M. More evident in Figure 4.8C, 5: State 0B (PIWIL4+) and 2: State

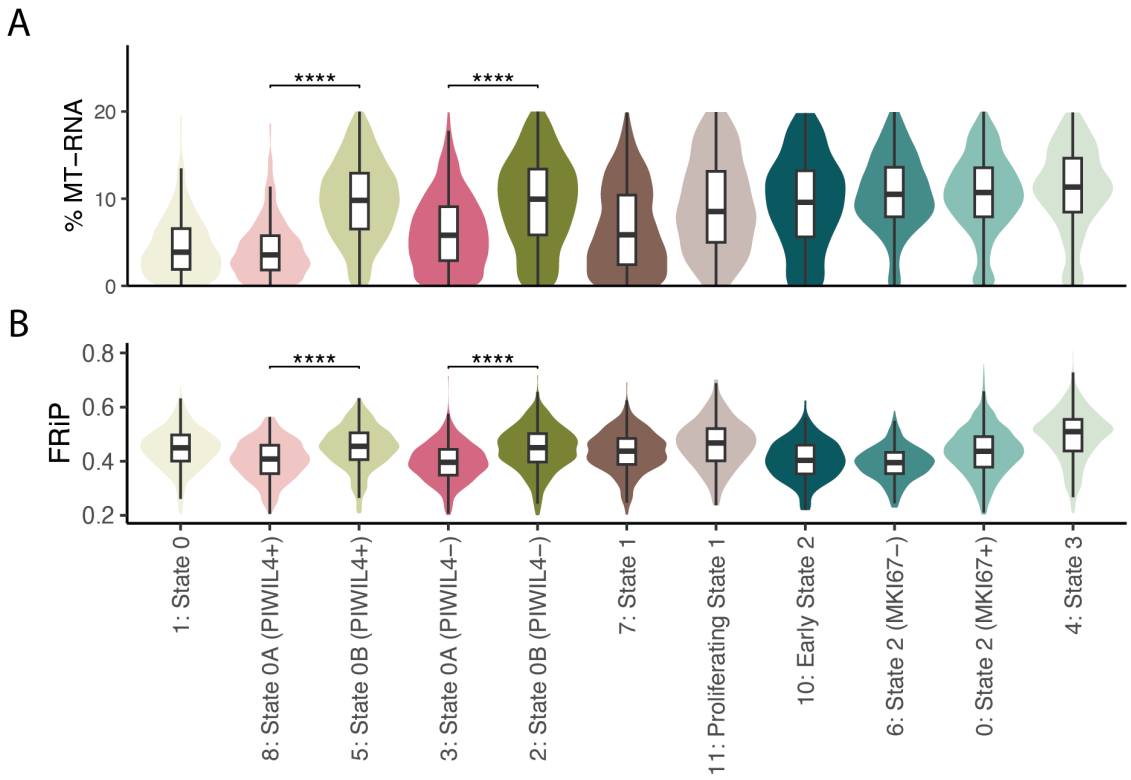



Figure 4.7: ***Exploration of confounding effects on the clustering of spermatogonia cell types in the adult human testis.*** **(A)** % of mitochondrial transcripts and **(B)** Fraction of reads in peaks (FRiP) for the spermatogonia subclusters. Significance for difference between State 0A and State 0B was calculated independently for the PIWIL4+ and PIWIL4- clusters using t-test and adjusted with Benjamini & Hochberg method. **** p.adj < 0.0001, *** 0.0001 < p.adj < 0.001, ** 0.001 < p.adj < 0.01, * 0.01 < p.adj < 0.05, ns 0.05 < p.adj < 1.

0B (PIWIL4-) have a slightly higher proportion of proliferative cells than 8: State 0A (PIWIL4+) and 3: State 0A (PIWIL4-) respectively. Ewen et al., 2013 also report a lack of correlation between the State 0A marker FGFR3 and the proliferating markers PCNA and MKI67 in immunohistochemistry. However, the UMAP obtained after regressing out the cell cycle scores from our dataset still retains the arrangement of State 0A and State 0B on either branches of clusters arising from 1: State 0 (Figure 4.8D). Furthermore, it appears that the % of mitochondrial reads (Figure 4.7A) and proportion of proliferating cells (Figure 4.8C) follow the same trend of enrichment in State 0B compared to State 0A clusters. Thus, the higher metabolic activity in State 0B possibly recapitulates cell proliferation rather than low quality cells.

Hence, while State 0A and State 0B show differences in the proportion of proliferative cells, cell cycle is not the key contributing factor that separates them into the two branches arising from 1: State 0 in the WNN UMAP.

## 4.6 Trajectory analysis of the spermatogonial compartment is inconclusive

One characteristic that could distinguish State 0A and State 0B is their position along the biological differentiation trajectory. The branched arrangement in the WNN

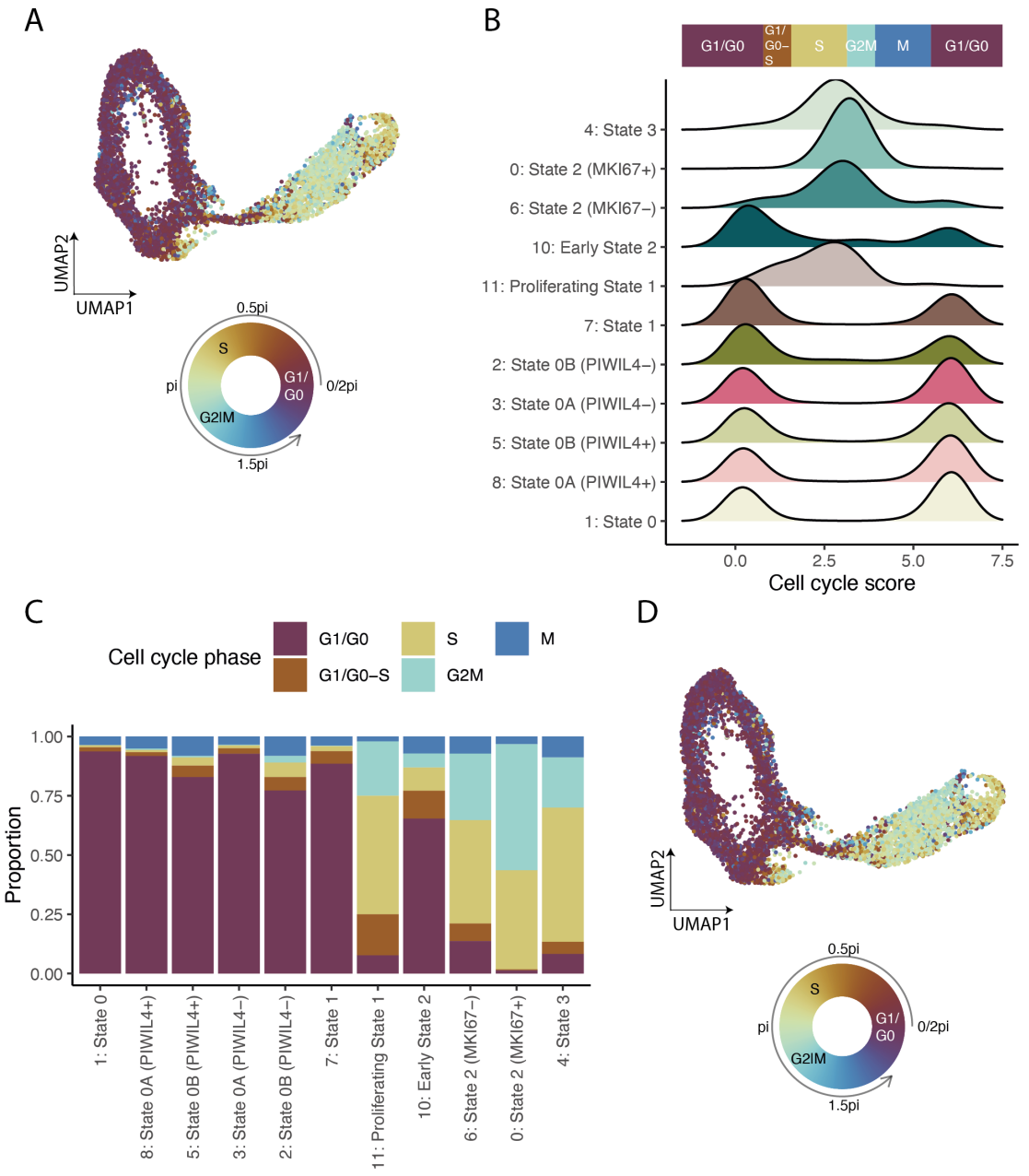



Figure 4.8: ***Cell cycle analysis with Tricycle.*** **(A)** Featureplot of Tricycle score on the WNN UMAP of spermatogonia subclusters. Tricycle assigns a continuous score where 0 or 2pi approximately corresponds to G1/G0, 0.5pi corresponds to beginning of S phase, pi corresponds to beginning of G2/M phase and 1.5 pi to middle of M phase. **(B)** Distribution of Tricycle score for each spermatogonia sub cluster. **(C)** Proportion of cells in the different cell cycle phases for the spermatogonia subclusters. Based on the approximate correspondence between Tricycle score and cell cycle phase mentioned in Zheng et al., 2022, cells with Tricycle score from 1.57 - 3.14 were assigned as "S" phase, from 3.14 - 3.925 were assigned to G2M" phase, 3.925 - 5.495 were assigned to "M" phase, score > 5.495 or < 0.785 were assigned to "G1/G0" phase and from 0.785 - 1/57 were assigned to “G1/G0-S" phase. **(D)** Featureplot of Tricycle score on the WNN UMAP of spermatogonia subclusters generated after regressing out tricycle score.

UMAP could be explained by some cells in 1: State 0 differentiating to 7: State 1 along State 0A and other cells differentiating along State 0B. This would point to the existence of heterogeneity in the 1: State 0 cluster, where individual cells within

the cluster have different propensities to differentiate along one branch or the other. Alternatively, cells at 1: State 0 could differentiate to 7: State 1 along one branch and cells at 7: State 1 could de-differentiate back to 1: State 0 along the other. The observed interconversion between a renewal-biased and a differentiation-primed state within the spermatogonial compartment in the mouse testis lends further support to this hypothesis (Nakagawa et al., 2021). To explore these possibilities, I performed trajectory analysis using two methods.

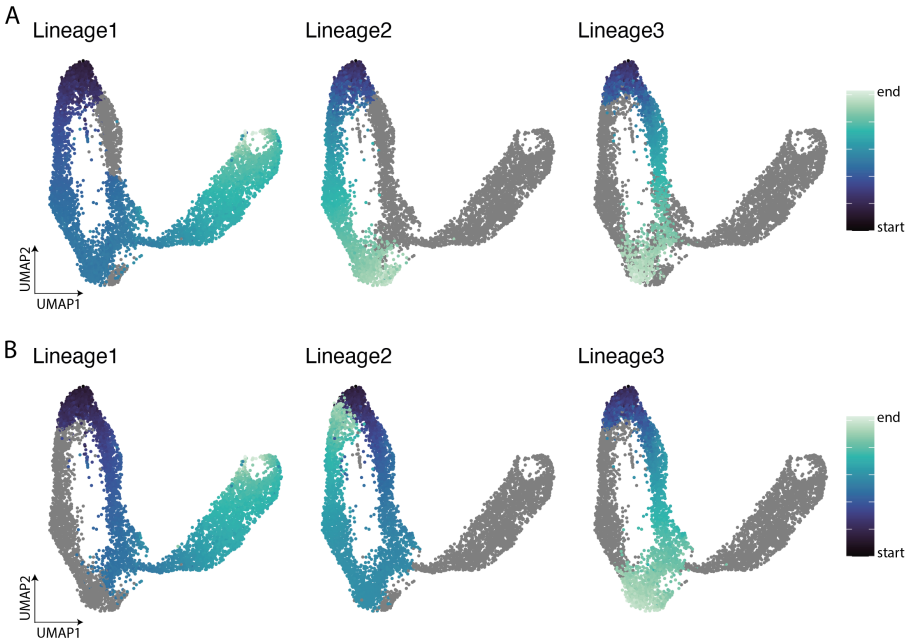



Figure 4.9: ***Trajectory inference of the spermatogonia subclusters with Slingshot.*** Trajectory inference using Slingshot with the cluster 1: State 0 assinged as the starting point and distance method specified as **(A)** "slingshot" (default), which computes the distance between cluster centres or **(B)** ”mnn”, which computes the distance between mutual nearest neighbour pairs.

First, I utilized the trajectory inference tool "Slingshot" (Street et al., 2018). Slingshot first identifies global lineages in a dataset based on the distance between cell clusters in a reduced dimension space such as the PCA. Then, it infers the position of a cell along these lineages. By default, Slingshot uses the cluster centroids to compute the distance between clusters. Using default distance method and assignment of "1: State 0" as the starting point produces three putative lineages of development (Figure 4.9A). Lineage 1 starts from the assigned starting point of 1: State 0 and continues via 8: State 0A (PIWIL4+), 3: State 0A (PIWIL4-), 7: State 1, 2: State 0B (PIWIL4-), 10: Early State 2, 6: State 2 (MKI67-) and 0: State 2 (MKI67+) to terminate at 4: State 3. Lineage 2 includes 8: State 0A (PIWIL4+), 3: State 0A (PIWIL4-), 7: State 1 and terminates at 11: Proliferating State 1. Lastly, Lineage 3 consists of cells transitioning from 1: State 0 via 5: State 0B (PIWIL4+) and 2: State 0B (PIWIL4-) to terminate at 7: State 1. Taken together, these lineages imply a scenario where State 0A and State 0B form two diverging branches along which cells can differentiate from 1: State 0 to 7: State 1. Termination of Lineage 2 and 3 at State 1 clusters suggests the existence of a temporary stationary point at 7: State 1 during the development of undifferentiated spermatogonia. 7: State 1 spermatogonia could require a second stimulus to undergo further differentiation. Furthermore, since 11: Proliferating State 1 does not form a part of the trajectory that undergoes complete differentiation until

4: State 3 (Lineage 1), it might primarily function as a source of replenishment of 7: State 1.

Curiously, in Lineage 1, Slingshot assigns cells on either side of the branch, specifically cells in 3: State 0A (PIWIL4-) and 2: State 0B (PIWIL4-) at similar positions along the trajectory. The default method of using cluster centroids to compute the cluster distances could result in such an inference. When clusters are present in an elongated shape, the centroids of parallel clusters are in closer proximity than the centroids of contiguous clusters. Thus, Slingshot probably inaccurately assigns a higher similarity score between State 0A (PIWIL4-) and State 0B (PIWIL4-), that reside on either branches, than between State 0A (PIWIL4-) and State 0A (PIWIL4+), that are adjacent to each other on the same branch. To account for such scenarios, Slingshot implements a method to compute the distance between the mutual nearest neighbours (MNN) of cluster pairs rather than their centroids. MNN pairs occur at the boundaries of two clusters. Thus, distance between MNN pairs will be shorter for clusters that are in contact with each other, than those that run parallel to each other in close proximity but never meet.

Implementing Slingshot with dist.method = "MNN" and 1: State 0 assigned as the starting point again produced three lineages (Figure 4.9B). Strikingly, Lineage 2 proposes cell transition from 1: State 0 via 5: State 0B (PIWIL4+) and 2: State 0B (PIWIL4-) and de-differentiating back to 1: State 0 via 7: State 1, 3: State 0A (PIWIL4-) and 8: State 0A (PIWIL4+). This suggests that State 0A and State 0B are similar cell states but have opposite cell fates along the developmental trajectory. Lineage 1 now represents cells undergoing complete differentiation from 1: State 0 to 4: State 3 via 5: State 0B (PIWIL4+) and 2: State 0B (PIWIL4-). Lastly, Lineage 3 is similar to the default implementation of Slingshot, where cells undergo differentiation from 1: State 0 via 5: State 0B (PIWIL4+) and 2: State 0B (PIWIL4-) and terminate at State 1. However, the lineage now terminates at 11: Proliferating State 1 rather than 7: State 1.

In our dataset, the lineages proposed by Slingshot vary dramatically based on the method of distance computation between the clusters. This variability makes it hard to interpret the observations. Besides, Slingshot solely depends on the similarity between clusters to determine the cell fate trajectory. Hence, I performed additional velocity analysis with a more sophisticated tool called MultiVelo, which is an extension of the RNA velocity framework (Li, Virgilio, et al., 2023).

Conceptually, RNA velocity utilizes the amount of unspliced and spliced transcript to estimate the future direction of a cell. By definition, RNA velocity represents the change in the spliced mRNA abundance along time. This change is determined by the balance between production and degradation/nuclear export of the spliced transcript. Under steady state conditions, the rate of the two processes are constant (Bergen, Lange, et al., 2020; La Manno et al., 2018). In Figure 4.10A, the slope of the diagonal represents the steady state equilibrium rate.

During induction of gene expression, there is an increase in the abundance of unspliced transcripts followed by the abundance of spliced transcripts. Thus, there is more transcript abundance than is expected from the equilibrium rate of the steady state. This increase in unspliced transcripts, followed by spliced transcripts, will be captured as an upward curve relative to the steady state diagonal as seen in Fig-

ure 4.10A. Similarly, during gene repression, there is a decrease in the abundance of unspliced transcripts followed by a reduction in spliced transcript abundance. Now, there is lesser unspliced and spliced transcripts than expected from steady state equilibrium rate. This drop in abundance will be captured as the downward curve in Figure 4.10A.

Thus, the velocity is calculated for individual genes, and the combination of velocities across multiple genes can be used to predict the future state of an individual cell. For every cell, these predictions are represented by the probability of transition to another cell. Thus, a developmental trajectory can be built based on the sequence of high-probability cell state transitions. When cell fate trajectory is represented as a Markov chain, which describes a sequence of transitions from one state to another where each transition occurs with a specific probability (Chain, 2017), cells can be traced to their state of origin or end. These states represent the root or end points of the trajectory. Thus, RNA velocity can further infer the directionality and the rate of cellular transitions.

MultiVelo expands on RNA velocity and utilizes the rate of chromatin opening or closing in addition to the rate of splicing dynamics to predict the fate of a cell.

MultiVelo proposes velocity vectors that suggest cell state transition from 4: State 3 to 1: State 0, which is opposite to the biological developmental trajectory (Figure 4.10B). Examination of the inferred terminal points of the trajectory reveals the assignment of differentiating spermatogonia as "root" cells and State 0 undifferentiated spermatogonia as "end" cells (Figure 4.10C). Constructing latent time based on the velocity vectors after correctly specifying 1: State 0 as "root" cells reveals a trajectory where spermatogonia transition from 1: State 0 to 4: State 3 with State 0A and State 0B placed at similar positions along the trajectory (Figure 4.10D). This could indicate that MultiVelo does not detect a difference between State 0A and State 0B in their developmental stages.

To explore the reason for the inverted velocity vectors in our dataset (Figure 4.10B), I examined the ratio of unspliced versus spliced counts for the top genes that contribute to parameter estimation of the model. Ideally, these genes should exhibit distinct upward and downward curvatures in the scatter arrangement of the cells for transcription induction and repression respectively, and thus inform the model whether an upregulation or a downregulation is occurring (Figure 4.10A). Unfortunately, the top 10 genes from our model do not display a pronounced "almond" shape of gene induction and repression in the cell scatter (Figure 4.10E). The solid black line represents the fit inferred by the model to explain the dynamics of the particular gene, based on the change in its ratio of unspliced vs spliced mRNA transcript across the individual cells. This is inaccurate for many of the top genes as described below.

Bergen, Soldatov, et al., 2021 provide an elegant review of the current challenges and limitations of velocity analyses, and explain how non-ideal gene splicing dynamics can lead to inaccurate RNA velocity estimations. In Figure 4.10E, *SCD5* exhibits multiple trajectories of change in the ratio of unspliced vs spliced transcripts. This indicates different production or degradation rates of spliced mRNA across the different clusters (I have assigned the different putative fits based on visual inspection as red and orange dashed lines). Hence, a single model is not sufficient to explain the overall dynamics of the gene. In such cases, the current implementation of velocity estimation from

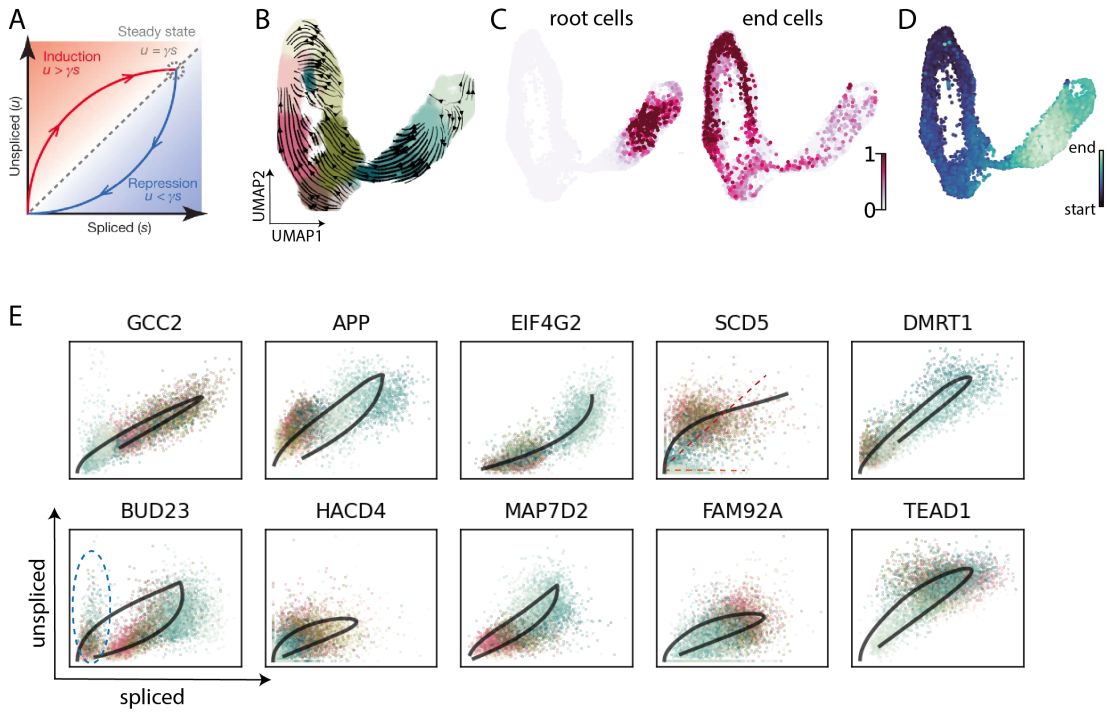



Figure 4.10: ***Trajectory inference of the spermatogonia subclusters with MultiVelo.*** **(A)** Schematic of a phase portrait indicating ideal changes in gene dynamics representing induction of gene transcription (upward curvature) and repression of gene transcription (downward curvature). *u* and *s* indicate abundance of unspliced and spliced transcripts and $\gamma$ represents the degradation/nuclear export rate. The image was adapted from La Manno et al., 2018. **(B)** Velocity embedding stream inferred by Multivelo on WNN UMAP of spermatogonia subclusters. **(C)** Transition probabilities of root or end cell states in the trajectory. **(D)** Latent time estimation by MultiVelo with a cell from the 1: State 0 cluster assigned as root cells. **(E)** Phase portraits or cell scatterplots representing spliced vs unspliced counts for the top 10 genes that contribute to model parameter learning. The cells are coloured by cluster assignment. The solid line represents the dynamics fit by Multivelo for the individual genes. *EIF4G2* and *SCD5* have been fit with repression and induction respectively. All other genes have been fit with both repression and induction phases. The blue dashed circle in *BUD23* and red and orange dashed lines in *SCD5* are manual annotations describing cells of interest and putative rates of change in the ratio of unspliced vs spliced transcript abundance as discussed in the text.

RNA alone offers an opportunity to fit these cell types independently when inferring gene dynamics (Bergen, Lange, et al., 2020). However, MultiVelo does not include this yet. *BUD23* is another example of a gene that exhibits differential kinetics. However, here, cells from the same cluster exhibit different rates of changes in the ratio unspliced to spliced transcripts. Thus, for example, cells within the blue dashed circle in *BUD23* are inferred to show increasing abundance of unspliced transcripts, followed by an increase in the abundance of spliced transcripts. However, these cells exhibit an increase in the unspliced transcript abundance without much change in spliced transcript abundance. *HACD4* is an example of a noisy gene, where the change in ratio of unspliced and spliced transcripts does not necessarily correlate with any cell state transition. *GCC2* shows a linear increase in the ratio of unspliced to spliced transcripts associated with steady state dynamics. Thus, it does not offer any curvatures associated with transcriptional induction or repression. Such cases could also lead to inverted velocity estimation due to the inaccurate assignment of downregulation fit for an induction phase of the gene and vice versa. Finally, *EI4FG2* is an example of a gene which shows a slight downward curvature representing transcription repres-

sion and could be informative for model learning. Taken together, most of the top genes selected by the model for parameter inference are not truly informative (such as *HACD4* and *GCC2*) or do not fit the current assumptions of the model (such as *SCD5* and *BUD23*). Thus, MultiVelo analysis cannot be used with our dataset to infer robust and accurate cell fate trajectories.

com

## 4.7 State 0A and State 0B exhibit differences in niche signalling

To inspect putative differences in the microenvironment of State 0A and State 0B spermatogonia, I analysed the cell-cell communication occurring in the testis. For this, I studied the ligand-receptor interactions (LRIs) between all pairs of testicular cells including the spermatogonia subclusters, germ cell population existing post spermatogonia development, and the supporting somatic cells. For the pairs of clusters, we examined the interaction pairs that are listed in the consensus database of LIANA (Dimitrov et al., 2022) as described in the Methods section.

The analysis reveals different number of interactions between pairs of cell type clusters (Figure 4.11A). The differences are more evident in a heatmap representation (Figure 4.11B) of the communication network (Figure 4.11A). As germ cell progress through spermatogenesis, the number of incoming or outgoing signals reduces from spermatogonia to spermatids (as indicated by the barplot representing the total incoming and outgoing signals for each cell type cluster in Figure 4.11B). Meiotic spermatocytes and haploid spermatids show the least number of interactions, possibly due to being located farther away from the basal membrane, on the other side of which many somatic cell types reside. Early stages of spermatogonia from 1: State 0 to 6: State 2 (MKI67-) display a high number of interactions with each other and with the somatic compartment. In general, the immune cell clusters, 20: macrophages and 22; T cells show minimal interactions with rest of the germ cells. The is possibly a consequence of the immune-privileged status of the adluminal compartment. Amongst all the cell type clusters, 1: State 0 and 0: Peritubular myoid cells constitute the receivers and senders of the maximum number of interactions respectively. The unique location of peritubular myoid cells, at the interface between the interstitial and the germ cell/Sertoli cell compartment possibly allows for interactions with most of the cell types within the testis.

Interestingly, meiotic spermatocytes and early spermatids, which show minimal number of interactions, exhibit the highest strength of incoming signals (Figure 4.12, left). Meiotic spermatocytes and especially post-meiotic round spermatids have been found to express the highest number of genes (Xia et al., 2020; Soumillon et al., 2013). This has been attributed to a burst of genome-wide transcription during meiosis and the spermatid retention of transcripts from previous stages (Johnson et al., 2011; Geremia et al., 1977). The diversity in transcriptome at these stages could impact inferences of cell-cell communication, wherein, a higher number of interactions could be inferred due to a higher number of expressed genes. However, it is unclear how the diversity in transcripts could affect the strength of interaction. 0: Peritubular

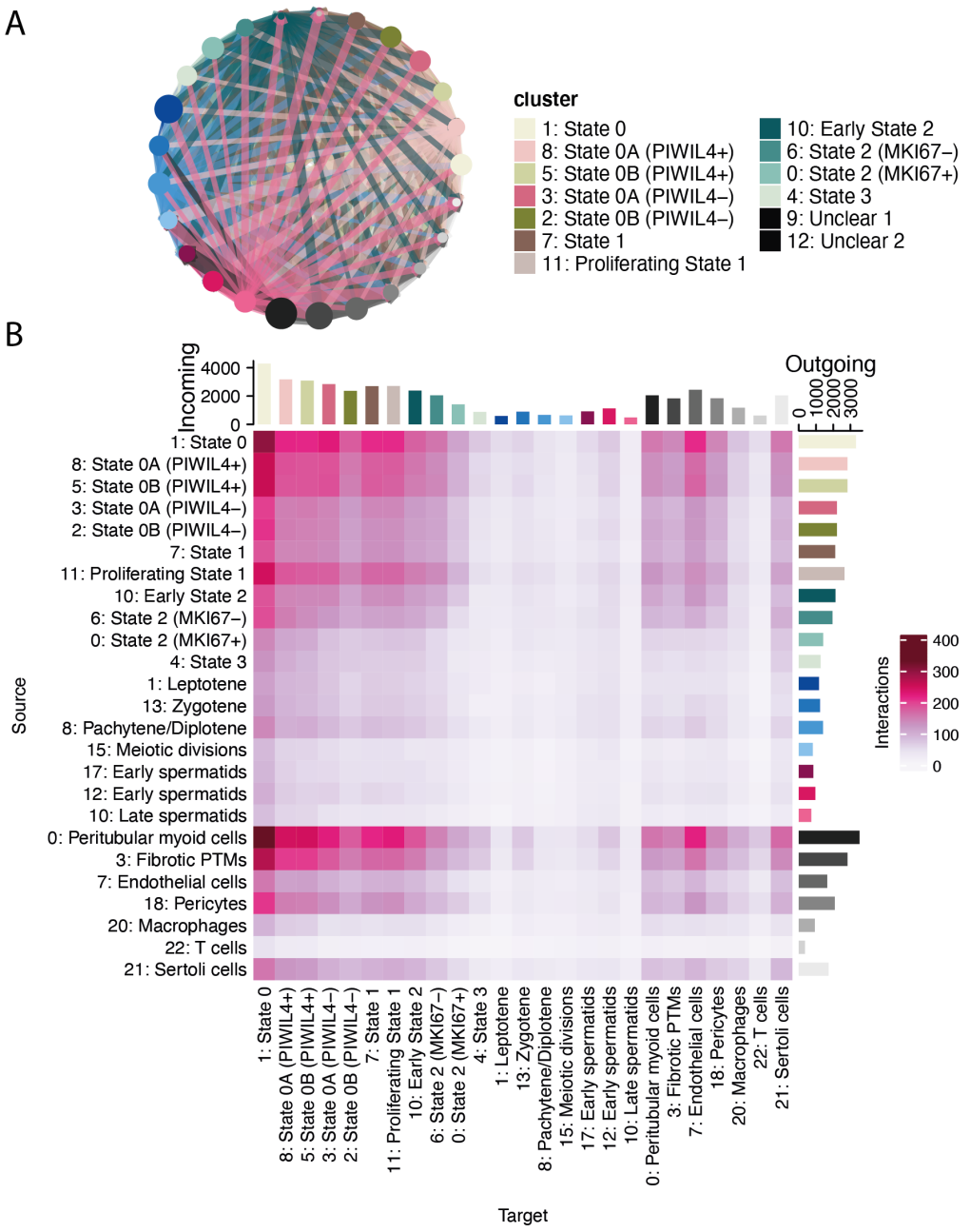



Figure 4.11: ***Number of ligand-receptor interactions in the adult human testis.*** The number of cell-cell interactions between pairs of clusters in the adult human testis. The LRIs included are significant with p adjusted value < 0.01. The ligand (on the source cell type cluster) and the receptor (on the target cell type cluster) of each interaction show an average expression of more than 0 in the respective clusters and are expressed in more than 10% of cells in the respective clusters. **(A)** Network representation where the size of the dot corresponding to the number of cells in the cluster and the colour corresponding to the cluster assignment. An arrow connects the source and target groups whose edge thickness represents the total number of interactions between the cluster pair. **(B)** Heatmap representation with the source clusters in rows and the target clusters in columns. The germ cells have been ordered based on the developmental trajectory followed by the somatic cells. The barplot at the top and to the right indicate the total number of incoming or outgoing signals for each cluster respectively.

myoid cells and 3: Fibrotic PTMs, apart from sending the most number of signals, also send the strongest signals (Figure 4.12, right). Notably, the two cell types send

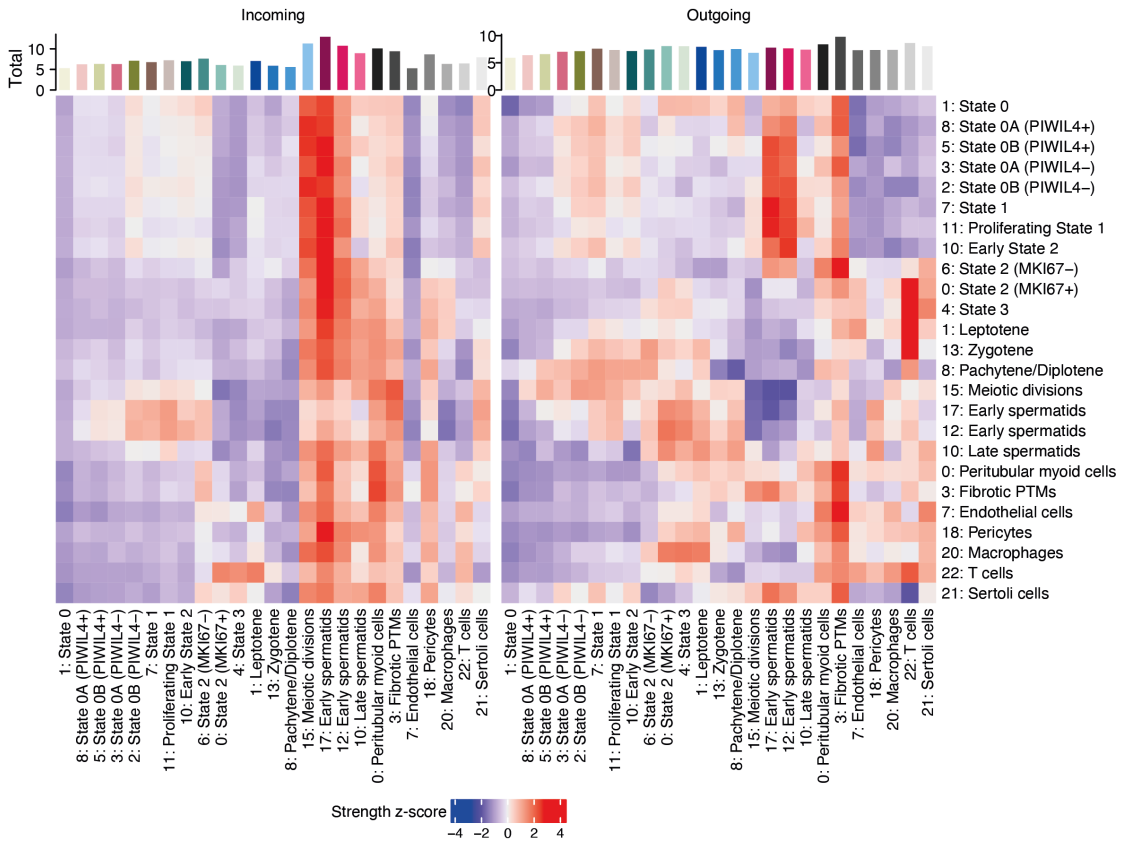



Figure 4.12: ***Strength of ligand-receptor interactions in the adult human testis.*** Heatmap of the strength z-score of interactions between pairs of clusters in the adult human testis for incoming signals to target clusters (left) and outgoing signals from source clusters (right). The germ cells have been ordered based on the developmental trajectory followed by the somatic cells. The barplot at the top indicate the total strength of incoming or outgoing signals for each cluster respectively. The LRIs included are significant with p adjusted value < 0.01. The ligand (on the source cluster) and the receptor (on the target cluster) of each interaction show an average expression of more than 0 in the respective clusters and are expressed in more than 10% of cells in the respective clusters.

slightly stronger signals State 0A than State 0B in both PIWIL4+ and PIWIL4- clusters. Within the immune compartment, T-cells receive relatively stronger signals from the somatic cells than the germ cells. However, the T-cells send strong signals to differentiating spermatogonia and early meiotic spermatocytes (i.e., cell type clusters from 0: State 0 (MKI67+) to 13: Zygotene). Lastly, Sertoli cells display differential communication between undifferentiated and differentiating spermatogonia, wherein it receives stronger signals from the former, but sends stronger signals to the latter.

When exploring the specific differences in signalling between State 0A and State 0B, I identified 588 and 282 significant interactions that were enriched in State 0A and State 0B (PIWIL4+) respectively, and 865 and 421 enriched in State 0A and State 0B (PIWIL4-) respectively. Calculation of the perturbation score, which represents the difference in LRI strength between State 0A and State 0B for a particular interaction with a specific source or target cell type, and the following filtering steps have been applied independently on the PIWIL4+ and the PIWIL4- clusters.

I selected the top interactions for plotting as follows: First, I removed autocrine signalling, and those interactions occurring within the same branch of the donut observed

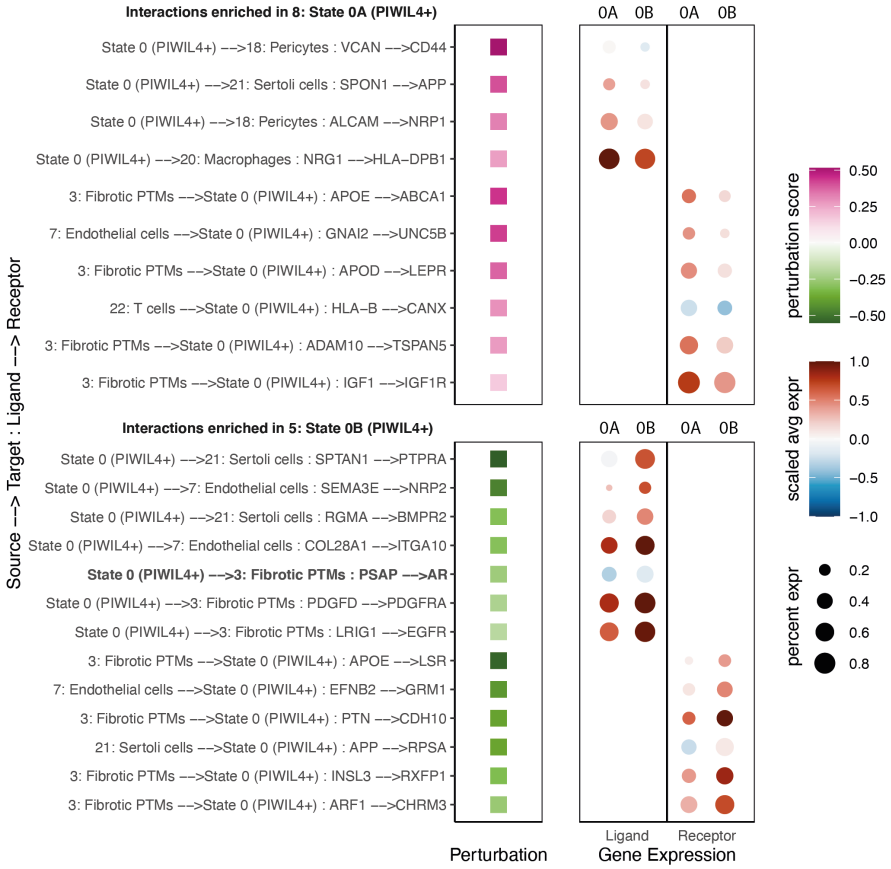



Figure 4.13: ***Differential LRI between 8: State 0A and 5: State 0B (PIWIL4+).*** Perturbation score, i.e., the difference in LRI strength between two conditions, and the gene expression of the corresponding ligand (in the source cluster) and the receptor (in the target cluster) in both State 0A and State 0B (PIWIL4+) clusters for the top LRI enriched in 8: State 0A (PIWIL4+) (top) or 5: State 0B (PIWIL4+) (bottom). Within each group, the interactions have been ordered based on originating from or being received by the State 0A or State 0B (PIWIL4+) cluster, and then on decreasing perturbation score. Colour of the dot indicates average of scaled expression and size indicates the % of cells in the cluster with detected expression. LRI that are discussed in the text are highlighted in bold.

in the WNN UMAP (for example, interactions between State 0A PIWIL4+ and State 0A PIWIL4-). Further, I only retained interactions between the germline and the soma. Then, for every LRI, I selected the source-target cell type pair that had the maximum perturbation score. This still left me with 27 and 31 LRIs for State 0A (PIWIL4+) and State 0B (PIWIL4+), and 116 and 39 LRIs for State 0A (PIWIL4-) and State 0B (PIWIL4-) respectively (data not shown). Hence, I performed two additional arbitrary filtering steps. First, for every receptor, I selected the ligand that showed the maximum perturbation score. Then, for every ligand, I selected the receptor with the maximum perturbation score.

It should be noted that all filters, especially the last two arbitrary ones, have been applied with the sole purpose of selecting a manageable number of interactions for plotting. A different set or sequence of filters will likely select slightly different LRIs for plotting. Significant interactions that have not been included in the following plots, are equally interesting for further exploration.

The current filtering results in 10 and 13 enriched LRIs in State 0A (PIWIL4+) and

State 0B (PIWIL4+) (Figure 4.13), and 21 and 18 LRIs in State 0A (PIWIL4-) and State 0B (PIWIL4-) respectively (Figure 4.14).

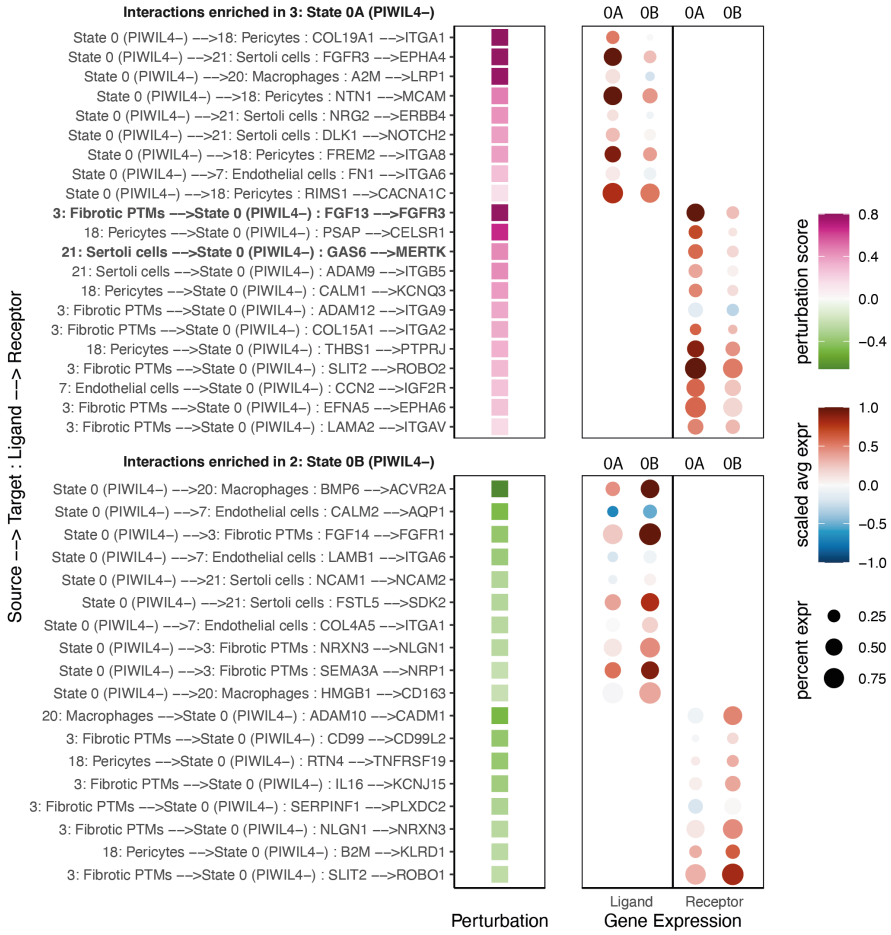



Figure 4.14: ***Differential LRI between 3: State 0A and 2: State 0B (PIWIL4-).*** Perturbation score and the gene expression of the corresponding ligand (in the source cluster) and the receptor (in the target cluster) in both State 0A and State 0B (PIWIL4+) clusters for the top LRI enriched in 3: State 0A (PIWIL4-) (top) or 2: State 0B (PIWIL4-) (bottom). Within each group, the interactions have been ordered based on originating from or being received by the State 0A or State 0B (PIWIL4-) cluster, and then on decreasing perturbation score. Colour of the dot indicates average of scaled expression and size indicates the % of cells in the cluster with detected expression. LRI that are discussed in the text are highlighted in bold.

Interestingly, State 0B (PIWIL4+) shows a stronger PSAP-AR (Ali et al., 2015) interaction with the target group of 3: Fibrotic PTMS than State 0A (PIWIL4+) (Figure 4.13). Androgen receptor (*AR*) is a well known player in male reproductive development and is expressed in multiple cell types within the testis. *AR* deletion in peritubular myoid cells has been reported to decrease sperm output (R.-S. Wang et al., 2009). Deletion in prosaposin (*PSAP*) also affects the development of multiple male reproductive organs (Morales and Badran, 2003) and the testes of mice carrying a homozygous mutation of *Psap* are smaller and exhibit impaired spemiogenesis (Morales, Zhao, et al., 2000). While the biological function of PSAP-AR interaction between State 0B (PIWIL4+) and F fibrotic PTMs is unclear, the requirement of both proteins for a wide range of male reproductive developmental processes presents an interesting avenue of exploration of their role within the spermatogonial compartment.

3: Fibrotic PTMs, through the ligand *FGF13*, interacts with the *FGFR3* receptor on 5: State 0A (PIWIL4-) (Figure 4.14). In the mouse testis, Kitadate et al., 2019 have elegantly demonstrated the regulation of balance between spermatogonial stem cell self-renewal and differentiation by "quorum sensing" of FGF ligands that are secreted by lymphatic endothelial cells.

Lastly, 3: State 0A (PIWIL4-), which has a higher expression of *MERTK*, is a stronger target for the GAS6 ligand originating from Sertoli cells, than 2: State 0B (PIWIL4-) (Figure 4.14). MERTK is a receptor tyrosine kinase that belongs to the TAM family that includes *TYROS3* and *AXL*. The TAM pathway is required for Sertoli cell phagocytosis and triple knockout of the TAM receptors in mouse leads to male infertility with an absence of mature sperm (Lu, Gore, et al., 1999; Lemke, 2013). Furthermore, mouse with specific homozygous knockout of the *Mertk* receptor exhibit age-related oligoasthenospermia. Apart from decreased phagocytic activity of Sertoli cells *in vitro*, these mice exhibit abnormal spermatogonia proliferation and differentiation (Shi et al., 2023). Thus, 3: State 0A (PIWIL4-), compared to 2: State 0B (PIWIL4-), could be subject to a stronger regulation of cell growth and development by the Sertoli cells through the TAM pathway.

**Contributions** The L-R interaction analysis was performed in collaboration with Paul-Georg Majev (Max Planck Institute for Molecular Medicine, Müenster), a shared PhD student between the Vaquerizas group and the group of Prof. Ralf Adams (Max Planck Institute for Molecular Medicine, Müenster). I pre-processed and provided Paul-Georg Majev with the adult human testis dataset containing the spermatogonia subclusters with rest of the germ and somatic cell types. Paul-Georg Majev implemented a methodology that he had optimized and utilized in his own PhD project to calculate the interaction strength and its significance for the set L-R pairs listed in the LIANA database. In this chapter, he computed the interaction strength and the significance of the interactions for the individual source-target pairs. He also computed the perturbation score and its significance when comparing State 0A vs State 0B. I performed the subsequent filtering, plotting and interpretation presented in this chapter. I also performed the CellChat analysis mentioned in the Discussion. I would like to thank Paul-Georg Majev for his collaboration. His implementation of LRI computation has been very beneficial for the analysis since it calculates a significance value when comparing conditions/datasets, which is not provided by other existing tools at the moment.

## 4.8 Discussion

In this chapter, I have conducted a detailed examination of the gene expression and associated chromatin accessibility landscape of the spermatogonial compartment present in the adult human testis. While an earlier study also examined the chromatin accessibility landscape of the adult human testis (Wu et al., 2022), the study only identified two clusters of spermatogonia and are thus unable to explore the landscape to the resolution achieved in this study. Apart from capturing previously known subtypes of spermatogonia, we identify eight new subtypes in our dataset including *PIWIL4+* and *PIWIL4-* subgroups of the undifferentiated spermatogonia States 0A and 0B, *MKI67+* and *MKI67-* subgroups of the differentiating spermatogonia State 2 and

an Early State 2 that is characterized by high *DMRT1* and low *KIT* expression. We also identify a subset of *GFRA1*+ State 1 cells that is proliferating, as seen by the co-expression of *MKI67*. This observation is in contrast to Guo, Grow, Mlcochova, et al., 2018, where the authors find almost no overlap between MKI67 and GFRA1, and thus propose a complete absence of a proliferative population before State 2. However, Di Persio, Saracino, et al., 2017 find a subset of MAGEA4+ spermatogonia that co-express GFRA1 and MKI67 in immunofluorescence staining of the human seminiferous tubules. 11: Proliferating State 1 is the smallest cluster in our dataset, and the limited number of cells probably did not form a separate cluster in earlier scRNA-seq studies of the human spermatogonial compartment.

The undifferentiated spermatogonia clusters in our dataset resemble a donut with State 0 at its apex and State 0A and State 0B forming either sides of the donut. While such an arrangement was observed by Sohni et al., 2019 upon re-clustering of the spermatogonia clusters, we already capture this arrangement in the WNN UMAP of the whole testis dataset in our study. Sohni et al., 2019 label these cell types as SSC1-B, SSC1-C and SSC1-A, and they correspond to State 0, State 0A and State 0B in our classification (Di Persio, Tekath, et al., 2021).

A key point to note in our discovery of new spermatogonia subgroups is that unsupervised clustering, in general, is capable of generating smaller and smaller subgroups depending on the parameters used. This issue is inherent in all single cell/nuclear studies where clustering forms a crucial, and many times, the first step of data analysis. Hence, it is very important to choose parameters that will result in an optimal number of meaningful clusters. In our study, we address this challenge by employing objective methods to determine the number of reduced dimensions to include in cluster optimization and by inspecting changes in cluster stability across multiple parameters. For the former, I used the R package *intrinsicDimension* (Johnsson et al., 2014) with maximum likelihood to estimate the intrinsic dimensionality of the snRNA-seq and snATAC-seq datasets. For the latter, I used the R package *clustree* (Zappia and Oshlack, 2018) to determine cluster stability at different cluster resolutions and different number of neighbourhood sizes. Recently, more robust frameworks that determine the significance of the clusters obtained, such as single-cell significance of hierarchical clustering (sc-SHC) (Grabski et al., 2023), have been published and can be implemented to increase the reliability of the clustering process. Furthermore, it is imperative to experimentally validate the new cell types obtained from computational analysis such as through immunofluorescence staining of human seminiferous tubules. We are currently coordinating with our collaborators at CeRA to perform the same on staged seminiferous tubules from healthy human testes for the subtype-specific markers we have identified.

Differential gene expression analysis reveals cell type-specific profiles and provides markers for the new subgroups identified. This includes genes such as *NRG1, TKTL1* and *MACROD2* that have been associated with male subfertility in different species (Zhang, Eto, et al., 2011; Talluri et al., 2022; Bischof et al., 2013). snATAC-seq data of the spermatogonial compartment suggests that the global genome accessibility decreases and then increases again as the cells differentiate from 1: State 0 to 4: State 3, with the least number of open peaks in 11: Proliferating State 1 and 10: Early State 2. A possible technical explanation for the fewer peaks identified in these two

clusters could be the limited number of cells resulting in lower coverage of snATAC-seq fragments (clusters 11: Proliferating State 1 and 10: Early State 2 are the two smallest clusters in the spermatogonia analysis). To tackle this, the number of open or closed peaks were identified after the calculation of average accessibility across all cells in the cluster. The peaks have further been identified after normalizing the data for sequencing depth. Moreover, these clusters exhibit similar, if not slightly higher, levels of average genome-wide accessibility compared to the other clusters.

We also identify a dynamic pattern of motif accessibility during spermatogonia differentiation, which is not captured in the bulk ATAC-seq comparison of human SSEA4+ undifferentiated and KIT+ differentiating spermatogonia in Guo, Grow, Mlcochova, et al., 2018. We detect cell type-specific enrichment of accessibility for motifs including NFYA/NFYB, CTCF/CTCFL which has earlier been observed in other cell types of the adult human testis (Wu et al., 2022). Additionally, we detect high activity of the CUX1 motif, which plays a crucial role in maintaining cell quiescence, in 1: State 0 cells (An et al., 2018). *PIWIL4*+ clusters 1: State 0, 8: State 0A (PIWIL4+) and 5: State 0B (PIWIL4+) have more motifs with higher accessibility compared to the other clusters. Yet, this does not necessarily occur with a very strong enrichment of transcription factor footprints in the corresponding motif.

Remarkably, 10: Early State 2, a cluster that lies at the transition point of undifferentiated and differentiating spermatogonia, presents a gene expression profile similar to the former and a motif accessibility profile similar to the latter. This probably reflects a rare transitioning population of cells that has just started differentiation. The identification of 10: Early State 2 in our dataset highlights the advantage of having joint measures of transcription and chromatin accessibility from the same single cells. Integrating independently produced sc/snRNA-seq and snATAC-seq datasets involves matching cells that show a correlation between the two measured modalities. Hence, a population like 10: Early State 2, which possesses transcription and accessibility profiles that each correlate with a different stage of development, might not be captured.

When examining the unique arrangement of State 0A and State 0B on either sides of a donut-shaped WNN UMAP with State 0 at its apex, we find that confounding effects such as cell stress or cell cycle are not the driving factors for this arrangement. Differences in measures of cell quality are inconsistent between State 0A and State 0B. State 0B shows a higher percentage of mitochondrial transcripts, indicating poor cell quality, but also a higher FRiP value, indicating high cell quality. Furthermore, State 0B includes a higher proportion of cells in proliferation. Thus, the increased percentage of mitochondrial transcripts possibly reflect the metabolic changes associated with cell division rather than a decreased cell quality. Furthermore, regressing out cell cycle score does not alter the donut-shaped arrangement of the spermatogonia subclusters. In general, undifferentiated spermatogonia, except 11: Proliferating State 1, are mainly in G0/G1 phase and the differentiating spermatogonia, are mainly in the proliferative (G2M, S and M) phase. This is in line with previous observations from scRNA-seq (Sohni et al., 2019) and immunofluorescence (Di Persio, Saracino, et al., 2017) studies. Interestingly, 10: Early State 2, a differentiating spermatogonia cluster, consists of cells primarily in G0/G1 phase and possibly reflects the "early-stage" nature of the cluster which is yet to commit to differentiation.

Trajectory analyses that explore the possible differences in developmental stages between State 0A and State 0B are also inconclusive.

In our dataset, Slingshot proposes drastically different trajectories based on the method of distance computation between the cell clusters. When using the centroids of the clusters to calculate the similarity between them, Slingshot proposes diverging trajectories from 1: State 0, where cells either transition along State 0A or State 0B during differentiation to 4: State 3. Additionally, it also predicts trajectories that terminate at 7: State 1 or 11: Proliferating State 1. This could indicate that State 1 cells form a stationary point during development, and require a second stimulation to undergo further differentiation. Alternatively, using MNN to determine cluster similarity predicts that cells developing from 1: State 0 to State 0B de-differentiate back to 1: State 0 via State 1 and State 0A. Cells that undergo complete differentiation produce a trajectory from 1: State 0 to 4: State 3 via State 0B. Earlier studies using similar methods to predict trajectory based on cluster similarity, such as Monocle (Qiu et al., 2017), capture similar trajectories that transition from State 0 to State 1 via State 0A or State 0B (Sohni et al., 2019).

A key limitation of methods such as Slingshot and Monocle is that trajectory inference based on similarity of clusters does not provide the directionality or rate of change of cell transition. Furthermore, Slingshot can only use one modality, i.e. either snRNA-seq or snATAC-seq to make predictions. In our study, we have implemented Slingshot with snRNA-seq data, and thus, the advantage of having the simultaneous measurement of chromatin accessibility is not fully availed. MultiVelo, on the other hand, employs both the change in splicing dynamics and chromatin accessibility to infer the directionality and rate of transition. In our dataset, MultiVelo predicts velocity vectors in opposite direction to known biology. Latent time inferred based on the velocity vectors with 1: State 0 specified as the starting point places State 0A and State 0B cells in similar position along the trajectory. A closer examination of the top genes that contribute to parameter estimation of the model reveals that these genes do not exhibit the required curvature to accurately distinguish induction and repression. Hence, the gene dynamics inferred by MultiVelo could lead to unreliable predictions. Lastly, both Slingshot and MultiVelo, much like most of the trajectory inference tools, still strongly rely on domain knowledge to manually determine the starting point of the trajectory.

An earlier study using RNA-only model to predict the velocity in healthy human spermatogonia development, has reported that only State 0B, and not State 0 or State 0A, show velocity streamlines towards State 1 (Di Persio, Tekath, et al., 2021). Furthermore, Guo, Grow, Mlcochova, et al., 2018 identify inverted velocity vectors from State 2 towards State 1 using a similar approach. Recently, Li, Virgilio, et al., 2023 have presented instances of biologically implausible backflows predicted by RNA-only models in embryonic mouse brain. These flows were re-inverted to follow the expected developmental trajectory after the inclusion of chromatin accessibility data. Thus, this raises uncertainty about the backflows reported in human spermatogonia development. Additionally, RNA velocity inference is most suitable for developmental processes that have a timescale comparable to that of the half-life of RNA molecules (La Manno et al., 2018). While the average half-life on mRNA is estimated to be 10-16 hours (Wang and Liu, 2022; Yang et al., 2003), human spermatogenesis occurs over a time period of $\sim$ 74 days (Heller, 1964). Thus, inferences of RNA velocity

may not provide accurate estimates in this system. Nonetheless, Nakagawa et al., 2021 have used more robust experimental approaches like lineage-tracing to show that *Gfra*+ spermatogonia in the mouse testis interconvert between a renewal-biased *Plvap*+ state and a differentiation-primed *Sox3*+ state to maintain a homeostasis in the spermatogonial compartment (Nakagawa et al., 2021). Hence, the possibility of de-differentiation in the human spermatogonial compartment requires further exploration such as through lineage-tracing experiments of *ex vivo* cultures of the human seminiferous tubule segments (Perrard et al., 2016).

Thus, the current analysis on the arrangement of State 0A and State 0B at distinct developmental time points, whether along differentiation or de-differentiation, remains inconclusive.

An exploration of the cell-cell communication in the testis reveals decreasing number of interactions within the germ cell compartment during progression of spermatogenesis. 1: State 0 are receivers of the maximum number of signals, and mainly from 0: Peritubular myoid cells. In fact, 0: Peritubular myoid cells send the maximum number of and the strongest signals across all cell types in the testis. The distinct positioning of the cells at the outer surface of the seminiferous tubules possibly enables its interaction with cells located both within the tubule and outside it. Interestingly, meiotic spermatocytes and spermatids are the receivers of the strongest signals. While these cells are known to possess the most diverse transcriptomes in the testis (Soumillon et al., 2013), it is unclear how diversity in transcriptome could lead to a higher strength of interaction and not number of interactions. Within the somatic cells, T-cells mostly show intra-somatic interactions. However, they exhibit very limited, but very strong, interactions with cells at mitosis-meiosis transition. Notably, 3: Fibrotic PTMs send slightly stronger signals to State 0A than State 0B.

A deeper examination of the differences in the interactions involving State 0A and State 0B reveals a stronger GAS6-MERTK signalling between Sertoli cells and State 0A (PIWIL4-). *Mertk* mutant mice exhibit abnormal spermatogonia dynamics (Shi et al., 2023) apart from Sertoli cell dysfunction (Lu, Gore, et al., 1999; Lemke, 2013). GAS6-MERTK interaction was also found to be enriched in spermatogonia in a recent spatial transcriptomic study of the human testis (Rajachandran et al., 2023). Similar to the study, we also identify various interactions involved in PDGF, EFRIN and Integrin signalling. While the study identified an enrichment of these signals in the spermatogonial compartment, we identify it as being differential between State 0A and State 0B. The spatial transcriptomics study grouped interactions into the different signalling pathways based on CellChatDB database. CellChatDB has the advantage of including information about multimeric complex formation and the effect of soluble and membrane-bound stimulatory and inhibitory co-factors (S. Jin et al., 2021). Hence, one would find fewer false positive interactions. I also captured the differential interactions such GAS6-MERTK and PERIOSTIN when performing LRI analysis with CellChat (data not shown). However, despite the obvious advantages, we decided to use LIANADB and perform the analysis as described in this chapter due to the following reasons. The consensus database from LIANA, despite lacking information about complexes and cofactors, includes interactions from 16 different resources including CellChatDB, CellPhoneDB and Omnipath, etc. Thus, LIANADB offers 4,701 LRIs, as opposed to the 1,939 LRIs available in CellChatDB for humans. Moreover, while the

current implementation of CellChat allows comparative studies across different conditions or datasets (such as the comparison between State 0A and State 0B), it does not provide a significance for the observed differences in interaction strength. Cell-cell communication is an upcoming field. Hence, there is currently no single tool that has been shown to outperform the others both in terms of the LRI database and the method of computation of interaction strength.

A key limitation of the analysis is that co-expression of a ligand-receptor pair at the transcriptomic level does not warrant their interaction at the protein level. This is a challenge intrinsic to all existing tools that infer cell-cell communication from sc/snRNA-seq data. Additionally, these approaches do not capture post-translational modifications that could alter the binding properties of a protein (Patwardhan et al., 2021; Geng, 2020). Hence, the interactions need to be stringently validated through techniques like co-immunoprecipitation or proximity labelling assays (Rao, Srinivas, et al., 2014; Zhou et al., 2020). This brings forth another challenge where cell-cell communication analyses capture many significant interactions, making it difficult to select specific L-R pairs for validation. The comprehensiveness of LRI databases may prove to be a double-edged sword unless complementary information such as co-factor interplay and subunit architecture is included as implemented in S. Jin et al., 2021 or Efremova et al., 2020. This is accompanied by the inconsistency of the top interactions selected based on the type and sequence of the filtering steps implemented. Cell-cell communication still relies heavily on literature to infer the importance of an L-R interaction or to deduce its effect on the gene expression of the target cell. To gain a better understanding of the downstream effects of L-R interaction, tools such as NicheNet have been developed that link ligands to target genes and model intracellular communication (Browaeys et al., 2020). Spatial transcriptomics, which includes information about the relative arrangement of cells in a tissue, is also advantageous since cell compartmentalization confers precision to L-R interactions (Iyer et al., 2023). Thus, integrating our multiome data with published spatial transcriptomics datasets (Rajachandran et al., 2023) will further aid in pursuing biologically functional L-R interactions in the adult human testis.

# Chapter 5

# Comparison of chromosome X- and Y-bearing spermatids

## 5.1 Introduction

### 5.1.1 Sex chromosomes in spermiogenesis

The process of spermatogenesis produces haploid cells called spermatozoa. These are the smallest cells in the human body and contain only one set of autosomes and the X or the Y sex chromosome (Alberts et al., 2002). Initially, the second division of meiosis produces haploid cells called spermatids. Spermatids then undergo extensive differentiation in a process called spermiogenesis to become mature spermatozoa (Gilbert and Barresi, 2000).

Spermiogenesis includes substantial remodelling of the DNA within the nucleus, whereby histones in the DNA are first replaced by transition proteins, and then by another class of proteins called protamines (O'Donnell, 2014). The histone-to-protamine transition enables tighter packaging of the genome to fit into the small sperm head, but also effectively renders the cell transcriptionally inactive. Furthermore, germ cells in the seminiferous tubule remain connected to each other through intracellular connections called cytoplasmic bridges that are $\sim$1 $\mu$m in diameter and arise from incomplete cytokinesis during spermatogenesis (Dym and Fawcett, 1971). Essentially, even though the spermatids are genetically haploid, these bridges enable them to be functionally diploid, since a gene product made in one cell can readily diffuse into the cytoplasm of its neighbours (Braun et al., 1989; Morales, Wu, et al., 1998). This also facilitates synchronous development of the cells (Huckins, 1971). During maturation, these bridges disintegrate and mature spermatids are released into the lumen as individual cells.

In humans and other therian mammals, primary sex determination, i.e., the development of gonads, is determined by the presence of an X- or a Y-chromosome (Gilbert and Barresi, 2000). *SRY* (Sex-determining region Y protein), a sex-determining gene on the Y-chromosome, becomes activated 6-8 weeks after fetus formation and inhibits the female anatomical structural growth in males. It also leads to the upregulation of SOX9 which produces the primary sex cords that develop into testicular seminiferous tubules (Larney et al., 2014).

Hence, an investigation of differences between spermatids bearing the X or the Y chromosome would help enrich for either of them and impact breeding programs in the livestock industry by aiding sex pre-selection.

### 5.1.2 Putative differences between chromosome X- and Y-bearing spermatids

Over the years, multiple studies have investigated the differences between chromosome X- and Y-bearing spermatids (referred to as X- and Y-spermatids respectively in rest of the chapter) in terms of their morphology (Carvalho et al., 2013; Shettles, 1960; Shettles, 1961), physiological properties such as motility, viability and stress response (You et al., 2018; Song et al., 2018; Sarkar et al., 1984; Umehara, Tsujita, et al., 2019) and in recent times, even their molecular characteristics such as transcriptomic and proteomic content (Chen, Yue, et al., 2014; Almiñana et al., 2014; De Canio et al., 2014; Scott et al., 2018; Chen, Zhu, et al., 2012). Many of these observations have been contradictory, either due to the use of limited and unspecific methods, or due to the highly contextual setting of the experiments. While Shettles, 1960; Shettles, 1961 proposed that the head of human X-spermatids was bigger than that of Y-spermatids when viewed under a phase contrast microscope, (Carvalho et al., 2013) found no such differences in bovine spermatid heads when using atomic force microscopy. Similarly, studies exposing spermatids to endocrine disruptors such as 2,3,7,8-tetrachlorodibenzo-p-dioxin (You et al., 2018), dibromochl oropropane and diazinon (Song et al., 2018) report a relatively decreased viability of Y-spermatids. However, in *in vitro* conditions mimicking the female reproductive tract, Y-spermatids survived longer than X-spermatids (Dominko and First, 1997; Holt and Fazeli, 2010). Rahman and Pang, 2020 provides an elegant review of key papers trying to decipher the differences between X- and Y-spermatids across multiple species.

More recent studies have also found X- and Y-spermatid-specific transcriptome and proteome profiles (Chen, Yue, et al., 2014; Almiñana et al., 2014; De Canio et al., 2014; Scott et al., 2018; Chen, Zhu, et al., 2012). However, a key challenge is the limited amount of RNA present in spermatozoa (0.01 - 0.1 pg in humans) compared to somatic cells (Rahman and Pang, 2020; Pessot et al., 1989; Krawetz, 2005; Jodar et al., 2013).

Single cell studies of the testis have collectively addressed many of these challenges including limited RNA in later stages of spermatogenesis, heterogeneity of the tissue and small size of biopsy tissue in humans. All previous transcriptomic studies using single cell technology to investigate the mammalian testis (Guo, Grow, Yi, et al., 2017; Guo, Grow, Mlcochova, et al., 2018; Guo, Nie, et al., 2020; Guo, Sosa, et al., 2021; Wang, Liu, et al., 2018; Hermann et al., 2018; Xia et al., 2020; Shami et al., 2020; Sohni et al., 2019; Di Persio, Tekath, et al., 2021), except one (Murat et al., 2023), capture both nuclear and cytoplasmic RNA. The presence of cytoplasmic RNA has made it unfeasible to distinguish X- and Y-spermatids, as it is a mixture of gene products from multiple cells facilitated by the exchange via cytoplasmic bridges. In 2023, during the preparation of this work, Murat et al., 2023 used snRNA-seq to computationally separate X- and Y-spermatids in humans and other organisms and detected genes that show preferential enrichment.

In this chapter, I use snATAC-seq reads to computationally classify spermatids into chromosome X- or chromosome Y-bearing. Using this classification, I find genes that are enriched in one or the other throughout the different stages of spermiogenesis. This includes an enrichment of *ZNF28*, a DNA-binding zinc-finger protein that has been proposed to regulate the silencing of the SINE-VNTR-Alu (SVA) retrotransposons (Fukuda, Makino, et al., 2022), in Y-spermatids.

## 5.2 snATAC-seq can be used to classify spermatids into chromosome X- and chromosome Y-bearing

X and Y chromosomes show differences in most of their sequence, except at the pseudoautosomal regions (PARs). PARs exhibit sequence homology between the sex chromosomes which allows their correct pairing and segregation during meiosis in males (Ciccodicola et al., 2000; Helena Mangs and Morris, 2007). To avoid ambiguity in the origin of reads arising from PAR regions, I removed genes or snATAC-seq fragments in peaks overlapping PAR regions from the dataset in all analysis done in this chapter.

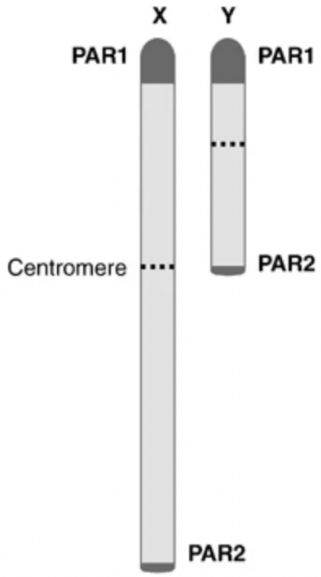



Figure 5.1: ***The human pseudoautosomal region (PAR).*** (Flaquer et al., 2008)

Since our multiome data consists of snRNA-seq and snATAC-seq libraries, I first compared the ability of both modalities to distinguish X- and Y-spermatids.

Figure 5.2A compares the fraction of transcripts arising from chromosome X vs chromosome Y for cells in the different germline clusters, revealing a clear separation into two populations from the meiotic divisions stage onwards, when sex chromosomes get segregated into different cells. In Figure 5.2B, the separation of cells into X- or Y-enriched populations becomes more pronounced when plotting them based on the fraction of snATAC-seq fragments in peaks arising from either sex chromosome. Hence, both modalities prove to be useful in separating X- and Y-spermatids, with snATAC-seq library being more promising than snRNA-seq.

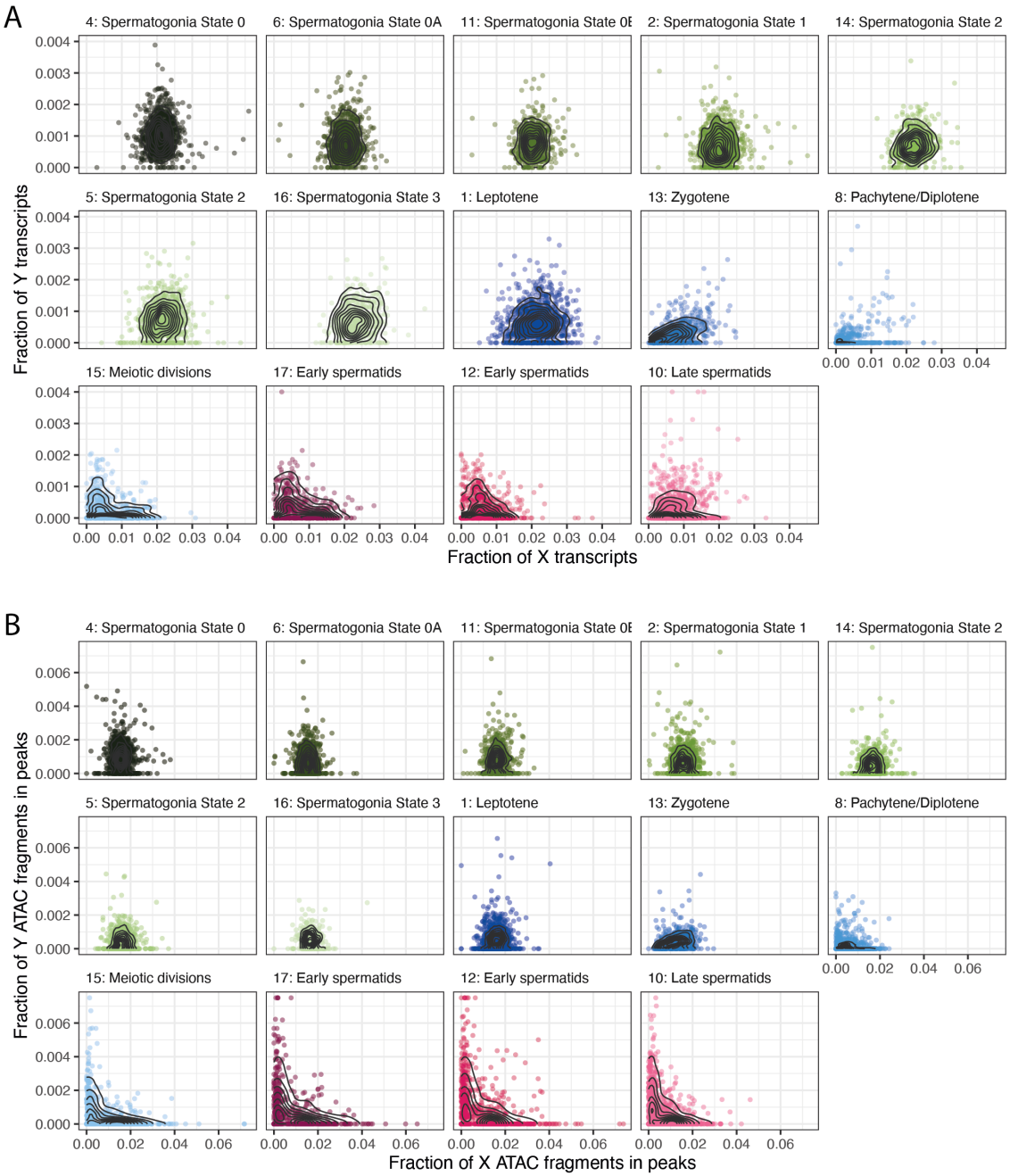



Figure 5.2: ***Distribution of reads from chromosome X or Y in human testicular germ cells.*** **(A)** Scatter plot of cells based on the fraction of transcripts arising from chromosome X or Y. **(B)** Scatter plot of cells based on the fraction of snATAC-seq fragments in peaks arising from chromosome X or Y. Reads overlapping PARs have been removed to avoid ambiguity in assigning their origin to either of the sex chromosomes.

To classify spermatids into chromosome X- or chromosome Y-bearing, I attempted two approaches - one based on snRNA-seq transcripts and the other based on snATAC-seq reads.

In the first approach, I implemented the methodology used in Murat et al., 2023. For each cell, I calculated the fraction of X transcripts. After combining cells from all spermatid clusters (early to late stage), I fitted a Gaussian Mixture Model with two components on fraction of X transcripts for each biological replicate. The resulting

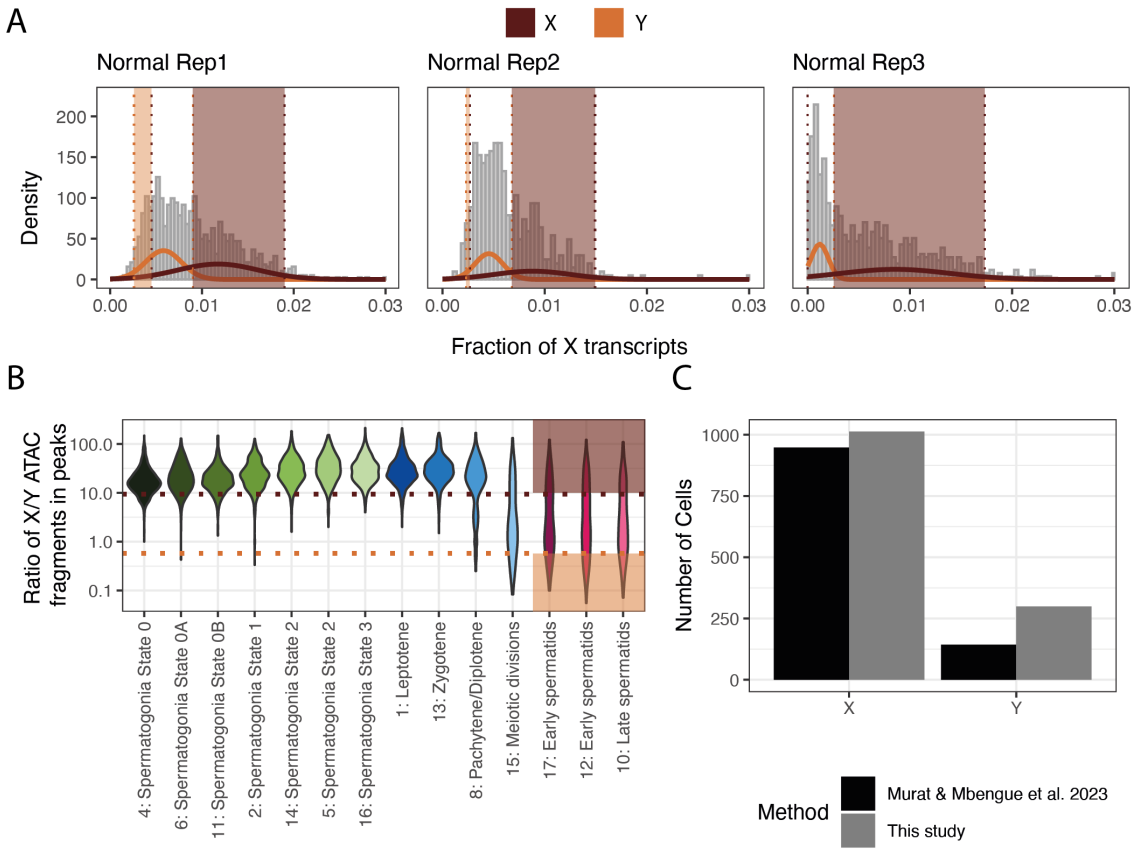



Figure 5.3: ***Classification of spermatids into chromosome X- or Y-bearing.*** **(A)** (Method: Murat et al., 2023) Histogram of fraction of X transcripts for spermatids from three biological replicates. The brown and orange lines depict the fitted curves for the bimodal distributions and the respective highlighted regions indicate populations selected as X- or Y-spermatids. **(B)** (Method: This study) Ratio of snATAC-seq fragments in peaks arising from chromosome X to chromosome Y for all germ cell clusters. Highlighted regions depict top 25% and bottom 25% of the spermatid clusters, representing X- (brown; higher ratio X/Y) and Y-spermatids (orange; lower ratio X/Y) respectively. **(C)** Number of X- or Y-spermatids obtained from the methods of classification used in Murat et al., 2023 and this study.

bimodal normal distributions with 95% confidence intervals (highlighted area in Figure 5.3A) were used to classify spermatids into chromosome X-bearing (higher levels of X transcripts) or chromosome Y-bearing (lower levels of X transcripts). Outlier and overlapping cells were classified into "neitherXnorY" and "bothXY" respectively and discarded from downstream analysis.

In the second approach, I calculated ratio of chromosome X/Y snATAC-seq fragments in peaks for each cell. At a cluster level, this distribution becomes bimodal from meiotic division onwards, similar to Figure 5.2B, representing the segregation of X and Y chromosomes into different cells. I first combined cells across biological replicates and different spermatid clusters (early to late stage). Then, of the cells that contained snATAC-seq fragments from both the X and the Y chromosome, I classified the top 25% of cells with the highest ratio of X/Y snATAC-seq fragments as X-bearing and the bottom 25% as Y-bearing cells (Figure 5.3B). Of the cells that contained snATAC-seq reads only from one of the sex chromosomes, I classified those as X-spermatids (or Y-spermatids) if they contained at least the minimum of chromosome X-associated (or chromosome Y-associated) snATAC-seq reads as a cell that was classified into

the same category based on its ratio of X/Y snATAC-seq reads. We capture a higher number of X- and Y-spermatids in the second approach than the first (Figure 5.3C). This difference is especially true for Y-spermatids, which is possibly due to the inclusion of reads on chromosome Y to distinguish the two populations in the second approach.

To validate the classification, I compared the percentage of transcripts and snATAC-seq fragments in peaks arising from the X or Y chromosome in these cells (Figure 5.4). I find that X-spermatids are enriched for transcripts and snATAC-seq fragments in peaks on the X chromosome compared to Y-spermatids and vice versa, for all spermatid clusters using both methods. While this enrichment is significant for all measures when classifying based on our approach, the enrichment for chromosome Y-associated transcripts or snATAC-seq fragments in peaks in Y-spermatids of cluster 10: Late spermatids is not significant when using the method proposed in Murat et al., 2023.

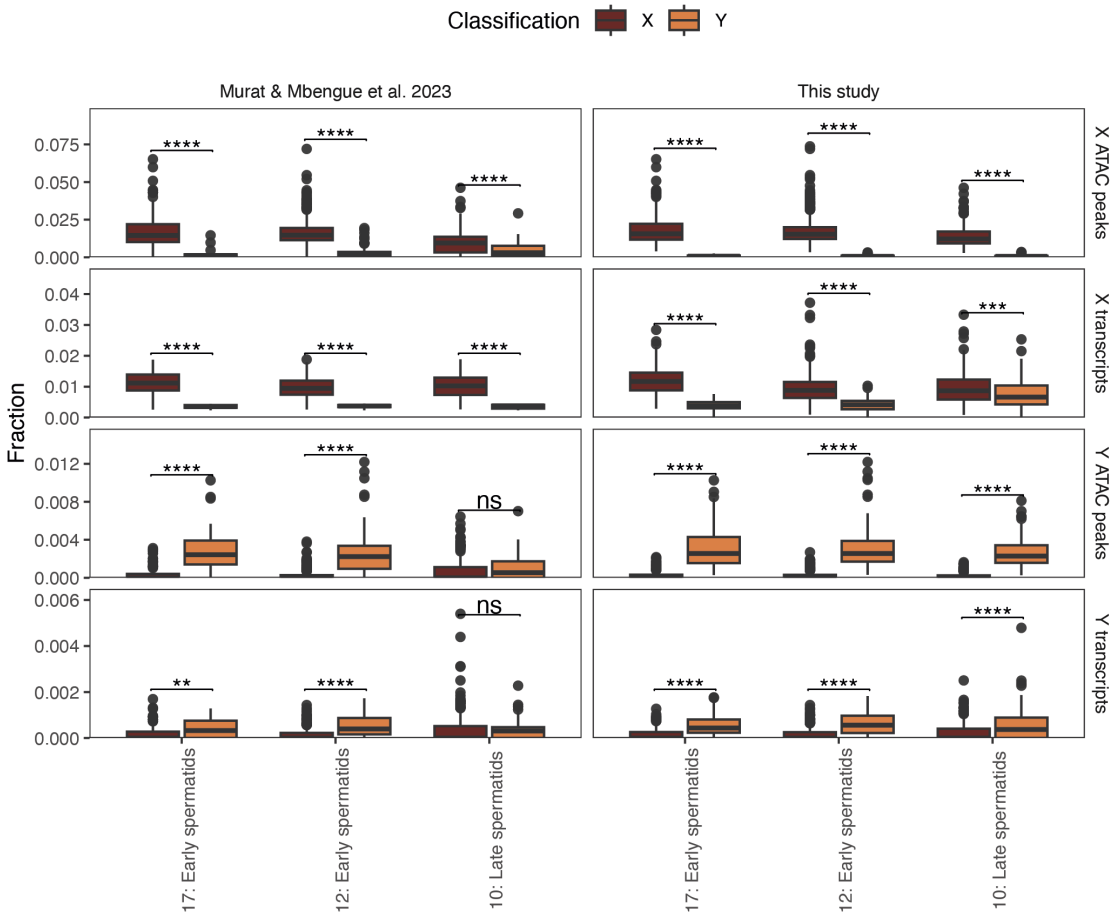



Figure 5.4: ***Fraction of transcripts or snATAC-seq fragments in peaks from the X or Y chromosome in cells classified as X- or Y-spermatids using method employed in Murat et al., 2023 or this study.*** Significance for difference between X- and Y-spermatids was calculated using t-test and adjusted with Benjamini & Hochberg method. **** p.adj < 0.0001, *** 0.0001 < p.adj < 0.001, ** 0.001 < p.adj < 0.01, * 0.01 < p.adj < 0.05, ns 0.05 < p.adj < 1

In summary, compared to fitting a bimodal Gaussian curve for fraction of X transcripts as done in Murat et al., 2023, using snATAC-seq reads from both the X and Y chro-

mosome captures a higher number of X- and Y-spermatids that show significant differences in the transcripts and snATAC-seq fragments in peaks from either of the sex chromosomes. Hence, I proceeded with the classification obtained from the second approach for all downstream analyses.

## 5.3 *ZNF28* is enriched in Early Stage Y-spermatids

To investigate possible differences in transcription between X- and Y-spermatids, I performed differential gene expression analysis between the two for each spermatid cluster.

In all clusters, X-spermatids show higher number of significantly enriched genes compared to Y-spermatids (Figure 5.5A). In general, we see that gene expression reduces in both X and Y-spermatids, as they differentiate from early to late stage (Figure 5.5B). This is in accordance with transcriptional silencing that accompanies spermatid maturation. Genes that are significantly enriched in X-spermatids are located on the X chromosome and thus possibly show enrichment solely due to the presence of the X chromosome. However, three out of four genes enriched in Y-spermatids are located on autosomes and could indicate a biological basis. ZNF28 is a DNA-binding zinc-finger protein located on chromosome 19, while FAM230C and FAM230F are long intergenic non-coding RNAs (lincRNAs) located on chromosome 13 and chromosome 22 respectively.

Since ZNF28 binds to DNA, I compared the accessibility in its motifs between X- and Y-spermatids. ZNF28 motif is indeed more accessible in early Y-spermatids compared to early X-spermatids (cluster 17: Early spermatids), concordant with gene expression (Figure 5.5C). Interestingly, the motif accessibility becomes comparable between the two at later stages (cluster 10: Late spermatids). This could potentially be explained by the sharing of ZNF28 protein from Y-spermatids to X-spermatids across the cytoplasmic bridges during their differentiation. However, when examining the footprint at ZNF28 motifs, the TF does not show visible differences in binding to its motifs between X- and Y-spermatids at any of the spermatid stages (Figure 5.5D). But this observation is not conclusive since the footprint signal is weak across all the spermatid clusters in both X- and Y-spermatids.

## 5.4 X- and Y-spermatids show differential accessibility at gene promoters

Spermatids experience progressive silencing of their genome as they differentiate to form mature sperm. To investigate possible differences in this silencing between X- and Y-spermatids, I performed differential accessibility analysis.

Overall, very few regions exhibit significant differential accessibility. In 17: Early spermatids, X- and Y-spermatids show 4 and 3 peaks that are preferentially accessible, in 12: Early spermatids, 19 and 6 peaks and in 10: Late spermatids, 7 and 9 peaks respectively.

The sparsity of the snATAC-seq data, the more condensed packaging of the DNA

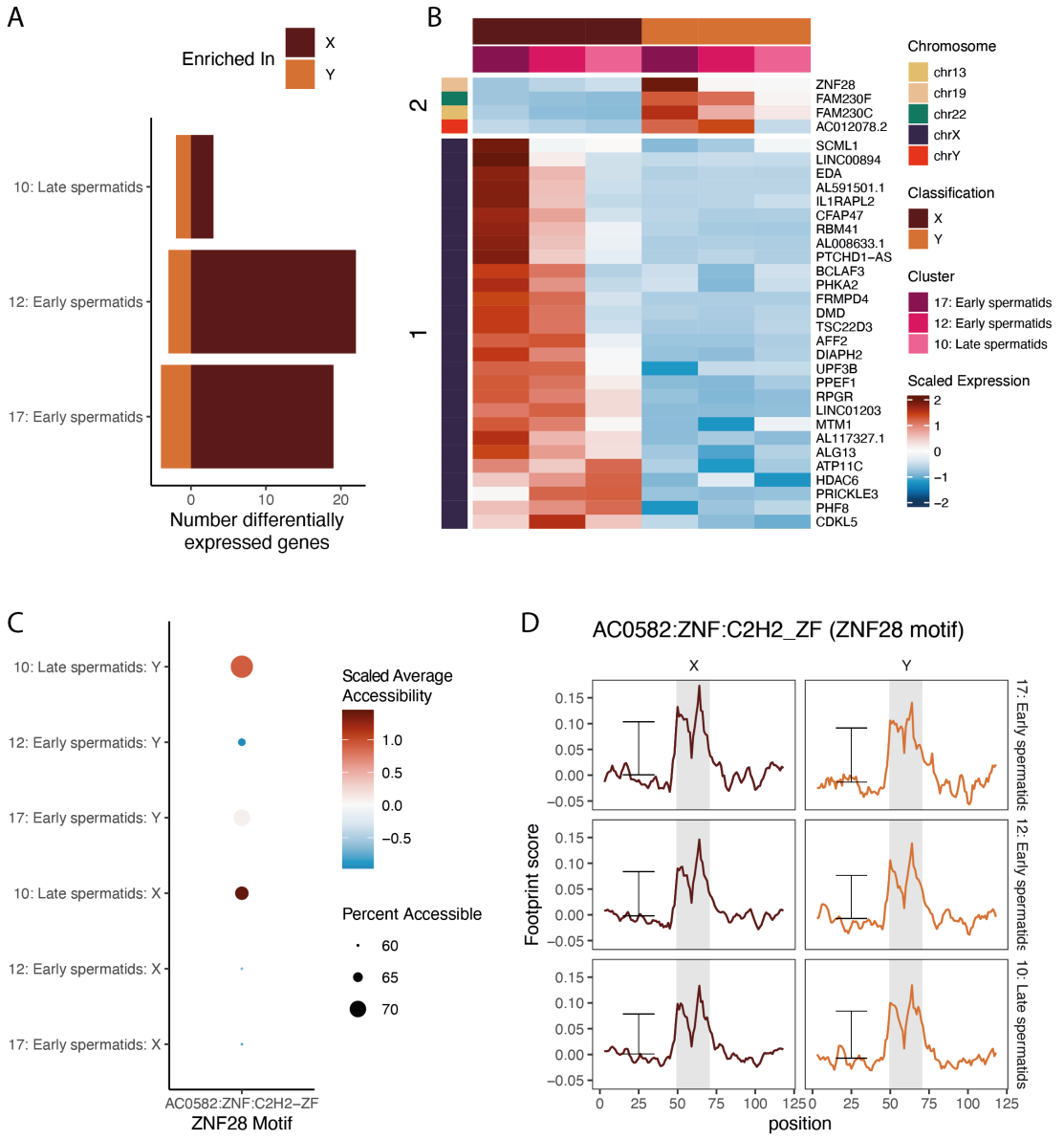



Figure 5.5: ***Differential gene expression between X- and Y-spermatids.*** **(A)** Number of differentially expressed genes between X- and Y-spermatids. **(B)** Heatmap of gene expression for differentially expressed genes. Clustering the heatmap based on gene expression produces 2 clusters - each consisting of genes enriched in X- (cluster 1) or Y-spermatids (cluster 2). **(C)** Motif accessibility of ZNF28 across all spermatid clusters. Colour indicates scaled average motif accessibility and size indicates the % of cells in the cluster with detected motif accessibility. **(D)** ZNF28 footprint across all spermatid clusters. Grey-shaded area represents the centre of the motif and the error bars depict the footprint depth or the difference in accessibility between centre and flank of the motif binding region.

leading to transcriptional silencing, and the limited number of spermatids classified into chromosome X- or Y-bearing preclude a more detailed and robust analysis of the chromatin landscape in haploid spermatids.

Figure 5.6A reveals that X-spermatids have higher number of accessible regions at the early stage (17: and 12: Early spermatids) and Y-spermatids at the later stage of differentiation (10: Late spermatids). All regions showing significantly higher accessibility in X-spermatids are present on chromosome X (Figure 5.6B) and overlap gene promoters (Figure 5.6C). However, Y-spermatids also show significantly higher

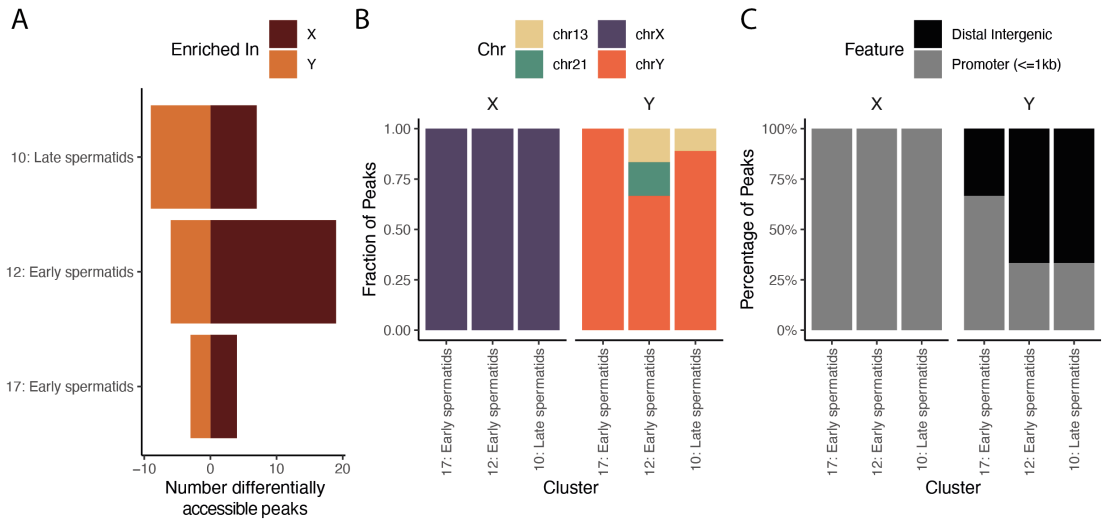



Figure 5.6: ***Genomic distribution of differentially accessible peaks between X- and Y-spermatids.*** **(A)** Number of differentially accessible peaks between X- and Y-spermatids for all spermatid clusters. Fraction of differentially accessible peaks from the X- or Y-spermatids **(B)** arising from different chromosomes or **(C)** overlapping different genomic features.

accessibility in the autosomes of chromosome 13 and 21. Y-spermatid-specific accessible regions also overlap distal intergenic regions apart from gene promoters.

In Figure 5.7A, I am showing examples of regions that have higher accessibility in X-spermatids for genes *CDKL5* (left panel) and *USP11* (right panel). For *CDKL5*, both promoter accessibility and gene expression are enriched in X-spermatids. However, for *USP11*, only the former and not the latter shows X-spermatid enrichment. The peak at *USP11* promoter does not correspond with expression of its neighbouring gene *CDK16* either. Hence, cell type-specific promoter accessibility does not necessarily correspond with cell type-specific gene expression.

Figure 5.7B depicts example regions showing Y-spermatid-specific accessibility. The left panel shows promoter accessibility and corresponding gene expression for *DDX3Y*. The right panel shows depicts intergenic accessibility within the gene body of *FAM230C*, which is also one of the lincRNAs that is enriched in Y-spermatids in Figure 5.6B. Since we find cases of both correspondence and the lack of, between promoter accessibility and gene expression within X- and Y-spermatids, I systematically analysed the relationship between them.

Firstly, within each cluster (early to late stage) for each type of spermatid (X- or Y-), I compared the differences in accessibility between regions that do or do not overlap a gene that is significantly enriched within the same population (Figure 5.8). Overall, I find only 13 peaks across all the developmental stages of X-spermatids that show this overlap. This includes the genes *SCML1, LINC00894, UPF3B, CFAP47, AL117327.1, RBM41, CDKL5, PHKA2, RPGR, MTM1, TSC22D3, PPEF1, HDAC6*. In Y-spermatids, no accessible regions overlap a significantly enriched gene promoter from the same population. In general, peaks that show this overlap have higher accessibility than peaks that do not. However, the analysis is limited by the minimal number of such cases.

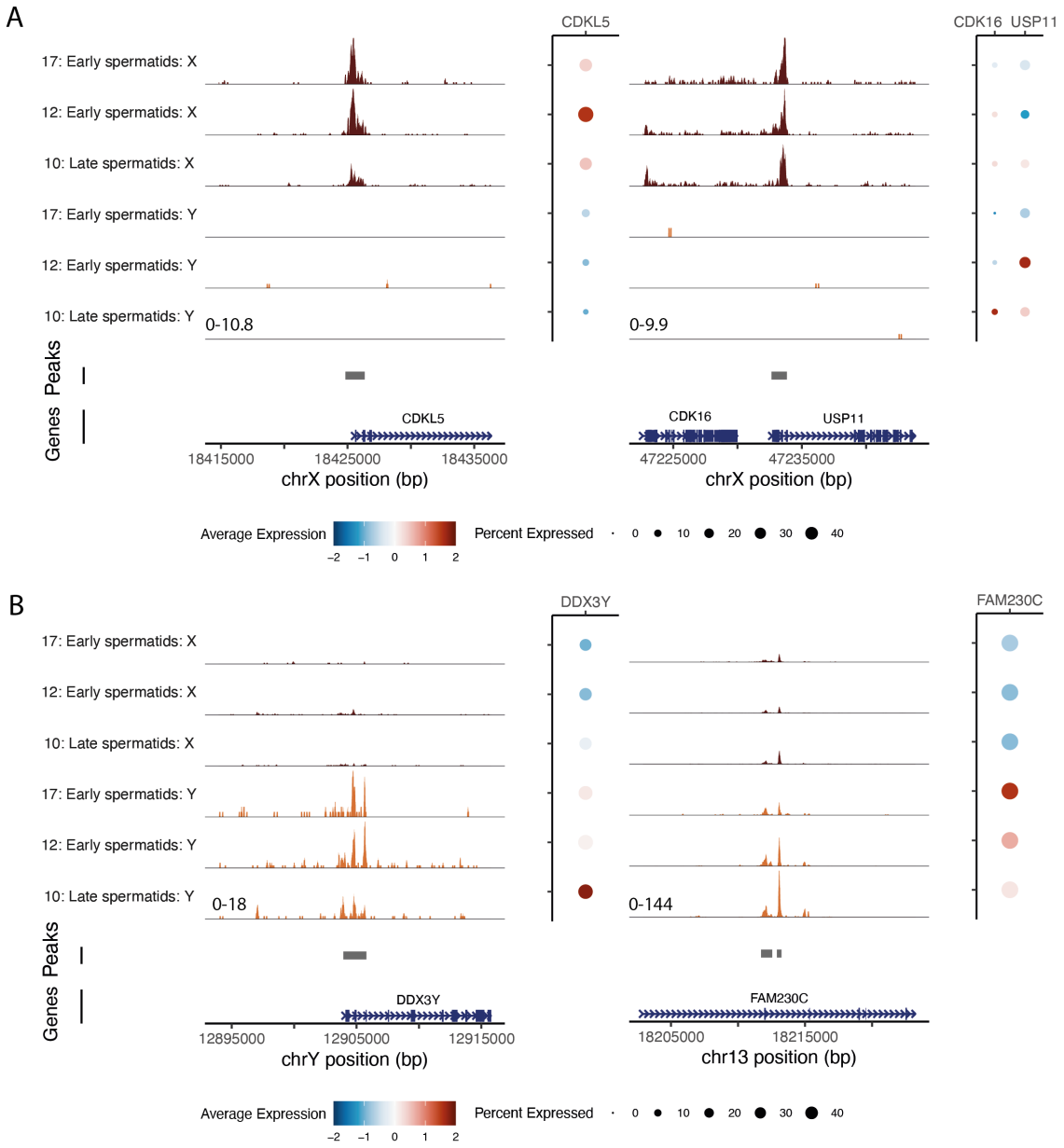



Figure 5.7: ***Genomic loci of X- or Y-spermatid-specific accessibility.*** Example regions of enriched accessibility in X-spermatids **(A)** or Y-spermatids **(B)**. Each sub-panel shows cluster-specific pseudobulk chromatin accessibility tracks on the left. Expression of the marker gene of interest in the cluster shown in the dot plots on the right, where colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression.

I also checked if the trend was the same the other way around - was there a difference in gene expression between genes that do or do not overlap a significantly accessible region in the same population? Again, I only capture 9 genes across all the developmental stages in X-spermatids. In Y-spermatids, only *DDX3Y* satisfies this criteria (Figure 5.9).

Hence, while there is a positive trend between promoter accessibility and gene expression (and vice-versa), the observation is severely limited due to very few genes or peaks showing an overlap. Due to global transcriptional and genomic silencing in spermatids, we capture very few genes or regions that are significantly enriched, es-

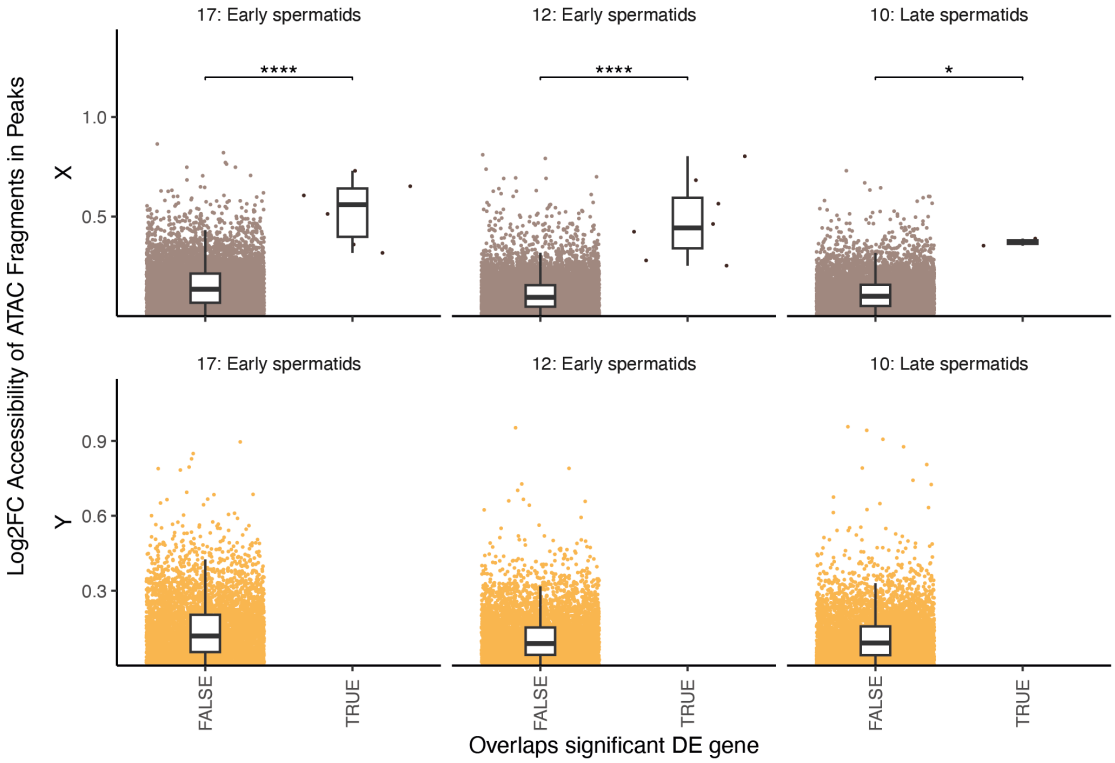



Figure 5.8: ***Accessibility of snATAC-seq fragments in peaks in X- and Y-spermatids categorized based on overlap with significantly enriched genes.*** The snATAC-seq fragments have been classified based on their overlap with a significantly enriched gene within the same population.

pecially on the Y-spermatids. The limited number of data points makes it challenging to obtain robust and significant associations.

## 5.5 ZNF:C2H2-ZF and NHLH/TFAP:bHLH motifs show higher accessibility in Late Y-spermatids

Lastly, to investigate possible differences in transcription factor activity between X- and Y-spermatids, I compared accessibility within known TF motifs. I found 2 motifs, AC0195:ZNF:C2H2-ZF and AC0078:NHLH/TFAP:bHLH, that are more accessible in Y-spermatids, specifically at the late stage (10: Late spermatids) (Figure 5.10).

However, the differences in TF footprinting within these motifs is negligible (Figure 5.11A). AC0195:ZNF:C2H2-ZF is predicted to be bound by ZNF621 and AC0078:NHLH/TFAP:bHLH by a list of TFs including *TFAP4, NHLH1, ASCL1, NHLH2, ASCL2, TCF12, ZBTB18 ZNF273, TCF3, MYF6, MYOG, MYOD1, BHLHA15, TCF7L2, BHLHE22, MYF5, FLI1*. None of the TFs show Y-spermatid-specific gene expression (Figure 5.11B). Interestingly, when examining the expression of these TFs across the whole testis, only *ZNF273* shows higher expression in spermatids along with the meiotic spermatocytes. A few TFs show high expression in spermatogonia including *ASCL1, NHLH2, TCF12, ZBTB18, TCF3, TCF7L2*. These TFs, except *ASCL1* and *NHLH2*, are also expressed in Sertoli cells. *FLI1* is only expressed in the somatic cell clusters of endothelial cells, macrophages and T cells. AC0195:ZNF:C2H2-ZF and AC0078:NHLH/TFAP:bHLH

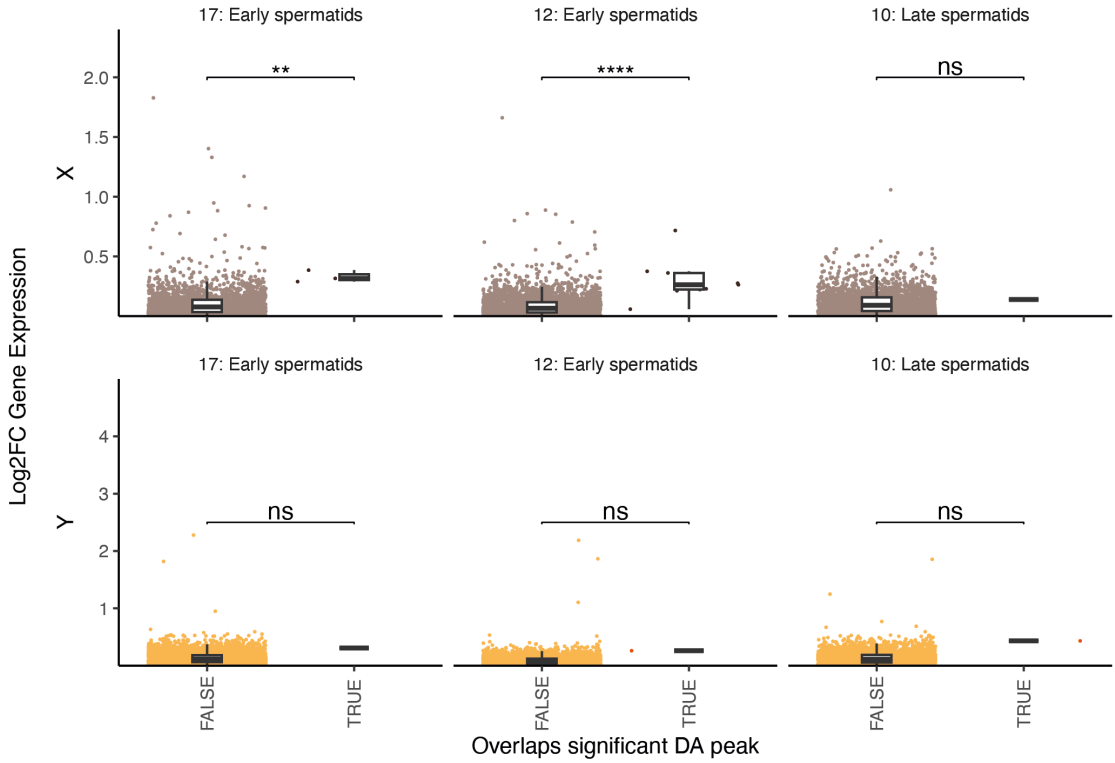



Figure 5.9: ***Expression of genes in X- and Y-spermatids categorized based on overlap with significantly accessible peaks.*** The genes have been classified based on their overlap with a significantly accessible peak within the same population.

also exhibit very high activity specifically in peritubular myoid cells and fibrotic PTMs (data not shown). Thus, the biological function of Y-spermatid-specific motif activity of AC0195:ZNF:C2H2-ZF and AC0078:NHLH/TFAP:bHLH is unclear.

## 5.6 Discussion

In this chapter, I have computationally categorized developing spermatids into chromosome X- and chromosome Y-bearing based on snATAC-seq data. Using a ratio of X/Y snATAC-seq fragments in peaks for each cell, I have identified cells as chromosome X-bearing (higher ratio) or Y-bearing (lower ratio). I further validated the classification by confirming that X-spermatids have significantly higher percentage of transcripts from chromosome X than Y-spermatids and vice versa. This classification has enabled a detailed characterization of the developing human X- and Y-spermatids at the genomic and transcriptomic level.

Previous scRNA-seq studies of the human testis have been unable to segregate spermatids into chromosome X- or Y-bearing, possibly due to mixing of cytoplasmic constituents via the intracellular bridges connecting germ cells across the tissue. However, in 2023, Murat et al., 2023 circumvented this issue by using snRNA-seq and successfully achieved the classification. Unfortunately, when implementing their method, I do not get an enrichment of chromosome Y-associated transcripts or snATAC-seq fragments in peaks in the resulting Y-spermatids at later stages (cluster 10: Late

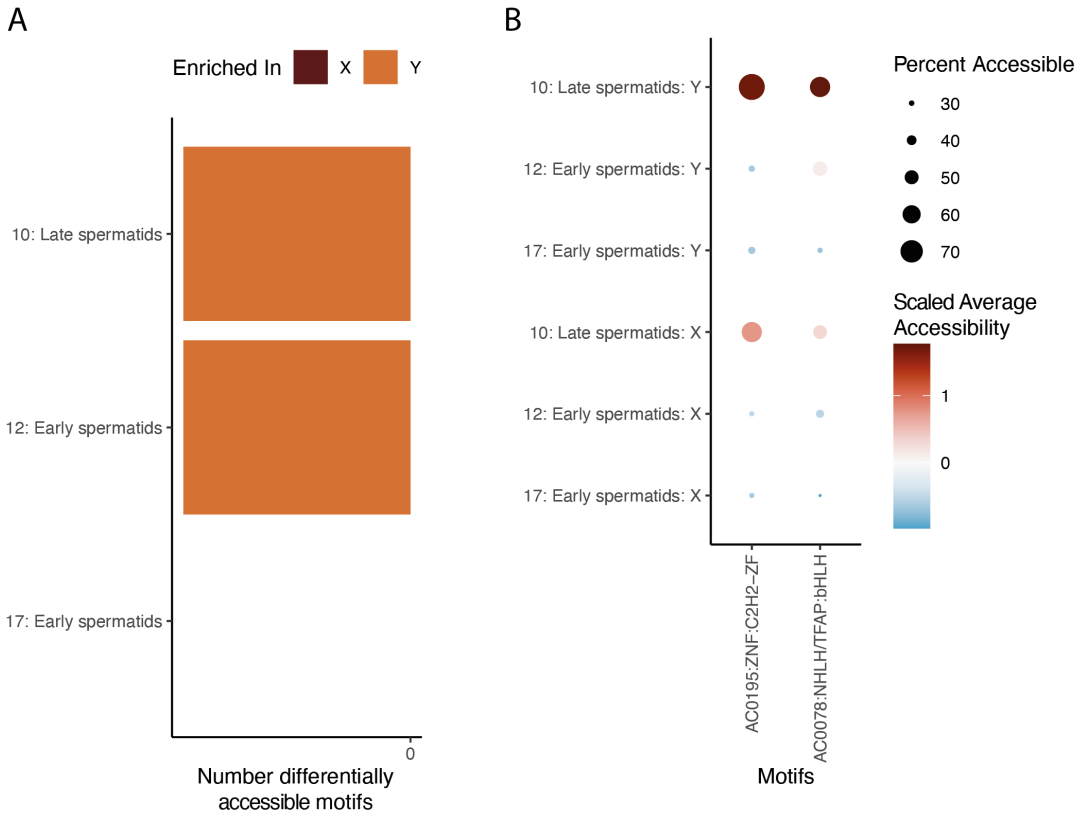



Figure 5.10: ***Differential motif accessibility between X- and Y-spermatids.*** **(A)** Number of differentially accessible motifs between X- and Y-spermatids in all spermatid clusters. **(B)** Average motif accessibility across all clusters of X- and Y-spermatids. Colour indicates scaled average motif accessibility and size indicates the % of cells in the cluster with detected motif accessibility.

spermatids). This could be due to a higher mixing of background RNA (termed "ambient RNA") across the single cells in our dataset, although ambient RNA is removed from each of our samples during the quality control steps (see Methods). A more concerning observation made during our analysis is the lack of clear separation of spermatids into solely chromosome X- or chromosome Y-bearing when focussing on the snATAC-seq reads. While the spermatid clusters indeed show preferential enrichment for snATAC-seq fragments in peaks arising from either of the sex chromosomes, most of the cells still exhibit a mixture of snATAC-seq reads from both sex chromosomes. While visualizing these pseudobulk snATAC-seq reads in IGV genome browser, I also see low background coverage on chromosome X in the Y-spermatids and vice versa (Figure 5.7). Although this is not biologically meaningful, it could indicate reads erroneously mapped because of similar sequence between the X and Y chromosome. However, only uniquely mapping reads with high mapping quality are retained in our analysis and transcripts or snATAC-seq fragments in peaks overlapping PARs, regions of similar sequence between the sex chromosomes, are also removed. Another possible explanation for the mixed X/Y snATAC-seq signal in haploid cells could be the leaking and subsequent mixing of DNA fragments across individual nuclei during the one-hour transposition step of the multiome library generation. Such mixing could possibly be avoided by fixing the nuclei before multiome library generation. Methanol-fixed nuclei did not give me successful snATAC-seq libraries (see Chapter "Generation and characterization of joint snATAC-seq + snRNA-seq multiome datasets of the adult human testis"), but alternative crosslinking agents such as Paraformaldehyde (PFA)

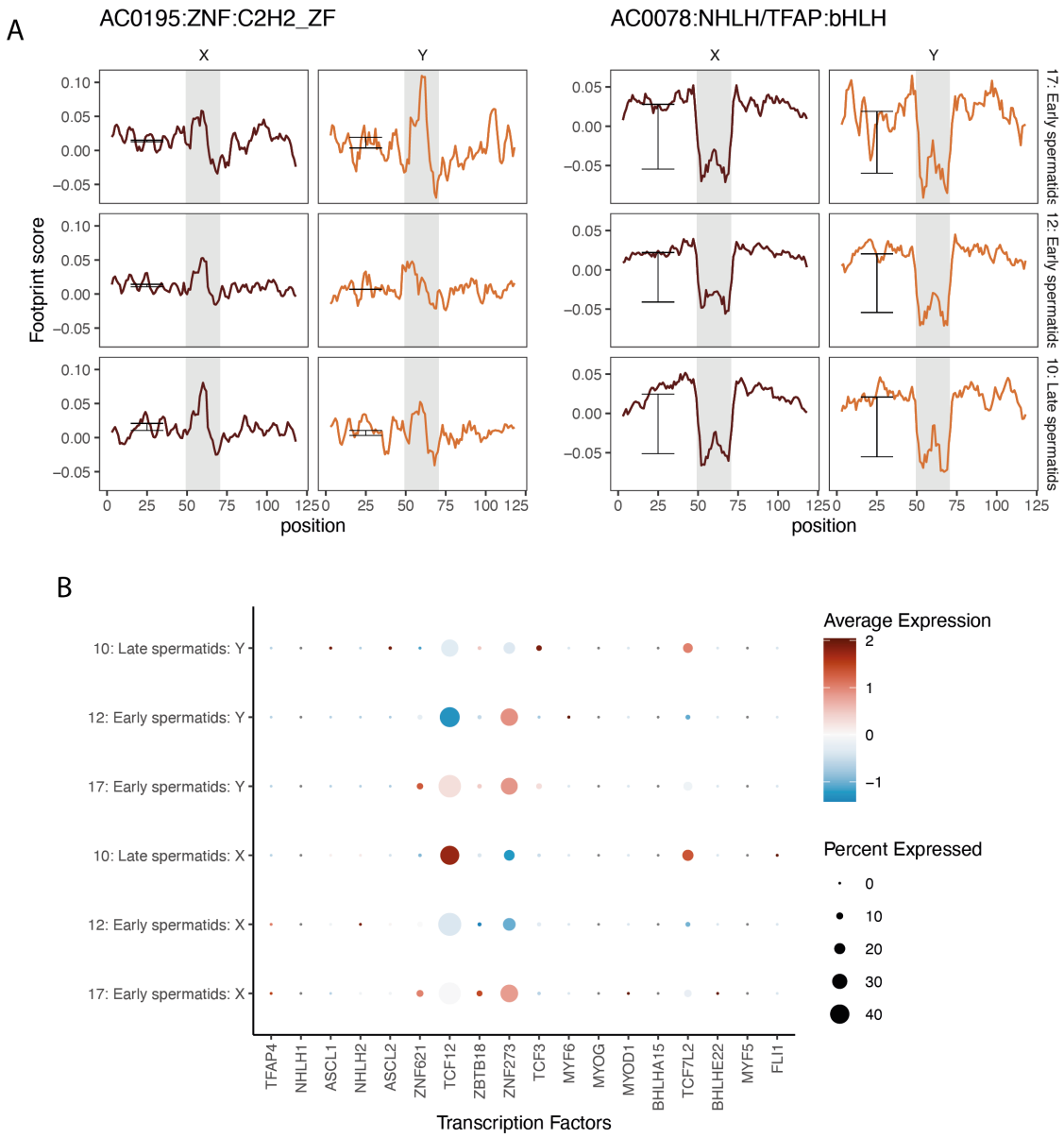



Figure 5.11: ***Footprint and gene expression of TFs predicted to bind to Y-spermatid-specific accessible motifs.*****(A)** Footprinting of Y-spermatid-specific accessible motifs across all clusters of X- and Y-spermatids. Grey-shaded area represents the centre of the motif and the error bars depict the footprint depth or the difference in accessibility between centre and flank of the motif binding region. **(B)** Gene expression of TFs that are predicted to bind to Y-spermatid-specific accessible motifs. Colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression.

could be explored. However, one would have to consider the impact of these chemicals on the downstream steps in the library generation protocol.

When comparing the X- and Y-spermatids obtained with our classification, overall, I observe limited number of differentially expressed genes or differentially accessible regions. This is in line with the silencing of the genome during spermatid maturation. While all genes enriched in X-spermatids are located on chromosome X, genes enriched in Y-spermatids include autosomal *FAM230C*, *FAM230F* and *ZNF28*.

FAM230C and FAM230F (also found enriched in Y-spermatids by Murat et al., 2023) are lincRNAs present on chromosome 13 and in the low copy repeat (LCR) regions of chromosome 22 respectively. FAM230C contains a DNA translocation element and the Translocation Breakpoint Type A (TBTA) sequence. Through gene duplication and subsequent modification, FAM230C has produced diverse long non-coding RNA (lncRNA) genes. Eight lincRNA genes that originate from the 3'half of the FAM230C sequence are found specifically in LCR22 of chromosome 22. Six out of these eight genes show expression exclusively in the testis (Delihas, 2018; Fagerberg et al., 2014). Furthermore, members of the FAM230 family, along with a newly found conserved long intergenic non-coding RNA (clincRNA) gene family, are found to form lincRNA gene and pseudogene families. Several uncharacterized proteins also carry the nucleotide sequence and open reading frame of FAM230. Thus, the FAM230 family of lincRNAs could play a role in the evolution of the testis transcriptome (Delihas, 2020).

ZNF28, on the other hand, is a DNA-binding Krüppel-associated box domain zinc-finger protein (KRAB-ZFP) that plays a putative role in the silencing of the hominoid-specific retrotransposon called SINE-VNTR-Alu (SVA) (Fukuda, Makino, et al., 2022). SVAs are a young family of retrotransposons that are active in humans and can cause genome instability, insertional mutagenesis, or transcriptional perturbation due to their transposition activity (Huang et al., 2012; Maksakova et al., 2013; Ostertag et al., 2003). Transposons can be silenced by mechanisms such as DNA methylation and histone H3 lysine 9 (H3K9) methylation (Fukuda and Shinkai, 2020; Goodier, 2016). Most of the retrotransposons are bound by specific KRAB-ZFPs that have co-evolved to bind them and recruit other repressive epigenetic modifiers (Schultz, Friedman, et al., 2001; Schultz, Ayyanathan, et al., 2002). During global DNA demethylation in human primordial germ cell (hPGCs) during embryonic development, a subset of retrotransposons, including a subgroup of SVAs termed SVA_A, are resistant to demethylation. Fukuda, Makino, et al., 2022 found that SVA_A harboured binding motifs for ZNF28 and ZNF257 in its variable-number tandem repeats (VNTRs). The authors also found that the enrichment of ZNF28/257 in SVA_A observed in HEK-293T cells correlates with SVA_A DNA methylation levels in hPGCs. In mature sperm, the authors also found differentially methylation regions within SVAs of individual humans that anti-correlate with their blood testosterone levels. In our analysis, we find a higher accessibility of the ZNF28 motif in early Y-spermatids compared to early X-spermatids, and this accessibility becomes comparable at later stages. This observation could indicate the production of ZNF28 in Y-spermatids and subsequent distribution to X-spermatids via cytoplasmic bridges so the protein can bind to and silence SVAs in both X- and Y-spermatids. However, ZNF28 footprint does not show the same trend. This hypothesis can be tested in two ways. First, we could perform differential transposon expression analysis between X- and Y-spermatids to check if the pattern of SVA silencing follows the pattern of ZNF28 motif accessibility. Secondly, one could also attempt to visualize the transport of ZNF28 from Y- to X-spermatids with microscopy, for instance, using time-lapse phase-contrast imaging of cytoplasmic bridges between spermatids *ex vivo* as done in Ventela et al., 2003 followed by simultaneous RNA-fluorescence in situ hybridization and Immunofluorescence (Meyer et al., 2017) to tag the X and Y chromosomes (to identify the respective spermatids) and the ZNF28 protein respectively.

When considering differential accessibility, we find very few peaks that are preferentially accessible in the X- or the Y-spermatids. Of the peaks that do show significant differential accessibility, those in X-spermatids are all present on chromosome X and overlap gene promoters. Accessible peaks in Y-spermatids occur on the autosomal chromosomes 13 and 21 and overlap distal intergenic regions in addition to occurring on chromosome Y and overlapping gene promoters. This includes accessibility in the gene body of *FAM230C*, which is one of the genes significantly enriched in Y-spermatids. There is a general trend for genes with more promoter accessibility to have higher gene expression and vice-versa, but this association is limited by few data points detected. While we observe other instances in our dataset where gene expression does not correlate with promoter accessibility (but rather with putative enhancer accessibility, see article included in the Appendix), this observation is especially understandable for spermatids - the genome of spermatids is mostly silent, however, the cells still retain transcripts produced in previous stages of differentiation (Geremia et al., 1977). Hence, gene promoter accessibility need not necessarily correlate with the transcript level of the corresponding gene.

Despite the overall silencing of the genome in late spermatids, we find 2 motifs that are more accessible in late Y-spermatids than late X-spermatids - AC0195:ZNF:C2H2-ZF and AC0078:NHLH/TFAP:bHLH. However, the TFs predicted to bind these motifs do not show Y-spermatid-specific gene expression or footprinting. Both motif activity and TF expression show a higher enrichment in other testicular cells. The motifs exhibit high activity specifically in the peritubular myoid cells and fibrotic PTMs while many of the TFs are expressed in the spermatogonia or in somatic cells. One could perform ChIP-seq for these TFs to further evaluate their binding targets and assess the possible differential binding between X- and Y-spermatids.

A key challenge with our analysis is the limited number of spermatids captured during generation of the snATAC-seq + snRNA-seq multiome libraries. As seen in the Chapter "Generation and characterization of joint snATAC-seq + snRNA-seq multiome datasets of the adult human testis", the multiome protocol captures fewer spermatids than scRNA-seq datasets due to the technicality of nuclei isolation. Even in scRNA-seq studies, Guo, Grow, Mlcochova, et al., 2018 report lesser and poorer quality of RNA transcripts associated with their spermatid clusters. Hence, implementing methods suggested in the Chapter "Generation and characterization of joint snATAC-seq + snRNA-seq multiome datasets of the adult human testis" to capture a higher number of spermatids could help detect more minute changes that are not captured in the current analysis. In summary, this chapter provides a detailed description of the differences in transcription and chromatin accessibility between the human X- and Y-spermatids and proposes a number of avenues such as differential TE expression analysis for further exploration.

# Chapter 6

# Dynamics of 3D genome organization across adult human spermatogenesis

## 6.1 Introduction

Spermatogenesis involves significant remodelling of the nuclear architecture, especially during meiosis and spermiogenesis. During meiosis, homologous chromosomes pair up and undergo recombination. This is accompanied by the simultaneous silencing of the sex chromosomes and their segregation into the XY body (O'Connor, 2008). During the subsequent stages of spermiogenesis, histones in the nucleus are replaced by protamines, leading to more condensed packaging of the DNA to fit into the small sperm head. This is also simultaneously accompanied by transcriptional silencing of the genome (Boussouar et al., 2008; Rousseaux et al., 2008).

3D genome studies of mouse and monkey spermatogenesis has captured dynamic reorganization of the chromatin. Meiotic architecture is accompanied by attenuation of TADs and sharpening of compartments. However, conventional 3D genome organization reappears in haploid spermatids (Alavattam et al., 2019; Wang, Wang, et al., 2019; Patel et al., 2019; Luo et al., 2020).

Surprisingly, the human mature sperm was found to lack TADs and this has been attributed to the absence of CTCF in them (Chen, Ke, et al., 2019). Recent studies have also been questioning the reappearance of conventional 3D genome organization observed in haploid spermatids due to possible contamination of cell free DNA (Yin et al., 2022). This is further supported by the lack of TADs observed in mouse mature sperm in one study that was hypothesized to have had increased accessibility to sperm DNA and reduced contamination of cell free DNA (Vara et al., 2019; Yin et al., 2022). Thus, observations of the 3D genome architecture in mature sperm is still inconclusive. This is especially true for human spermatogenesis, where there is additionally a lack of 3D genome maps of all the preceding developmental stages leading up to mature sperm.

In this chapter, we have established a FACS strategy to obtain the major germ cell populations across adult human spermatogenesis including undifferentiated spermato-

gonia, differentiating spermatogonia, meiotic spermatocytes and haploid spermatids. Utilizing this, I have generated high resolution 3D genome maps of each of these stages using a novel low input Micro-C protocol optimized in our lab. The resulting Micro-C maps show evidence of chromosomal territories in all the stages of spermatogenesis. They capture meiotic chromatin condensation in spermatocytes and reveal dynamic domain boundaries in spermatogonia that are lost during meiosis and regained in spermatids. In addition, the X chromosome in spermatocytes, unlike the autosomes, shows a specific pattern of contacts between its two ends and possibly represents the folding of the sex chromosomes into the XY body. Interestingly, snATAC-seq + snRNA-seq multiome data of the adult human testis reveals decreasing CTCF footprint during human spermatogenesis. In summary, our datasets can provide a detailed map of the 3D genome re-organization across human spermatogenesis.

## 6.2 Establishment of novel 5-colour FACS strategy to enrich for key developmental stages of human spermatogenesis

To dissect the changes in the 3D genome organization along human spermatogenesis, I, along with Sara Di Persio (Centrum ür Reproduktionsmedizin und Andrologie (CeRA), Muenster, Germany), established a FACS strategy to obtain germ cells from key developmental stages including undifferentiated spermatogonia (UD-SPG), differentiating spermatogonia (Diff-SPG), meiotic primary spermatocytes (SPC-I) and haploid spermatids (SPT).

Figure 6.1A depicts the major germ cell types present across human sperm development, along with their chromatid content. A subset of the undifferentiated spermatogonial stem cells (UD-SPG) (2C chromatids content) in the seminiferous tubules of the testis commit to differentiation (Diff-SPG). Differentiating spermatogonia subsequently double their DNA to become 4C primary spermatocytes (SPC-I). These cells undergo two consecutive meiotic reduction divisions to give rise to 2C secondary spermatocytes (SPC-II) and 1C spermatids (SPTs). In our FACS strategy (Figure 6.1B), we take advantage of this difference in chromatid content and use Hoechst 33342, a DNA stain, to enrich for pure populations of 4C primary spermatocytes, 1C early/late spermatids and a mixed population of 2C undifferentiated and differentiating spermatogonia, secondary spermatocytes and somatic cells. From the mixed 2C population, we first enrich for the spermatogonia with the pan-spermatogonia marker MAGEA4. From the spermatogonia population, we finally enrich for undifferentiated and differentiating spermatogonia with the markers UTF1 and KIT/DMRT1 respectively. While KIT is a membrane marker, MAGEA4 is a cytoplasmic marker and UTF1 and DMRT1 are nuclear markers. In order for the intracellular antibodies to enter the cells, we fixed and permeabilized the cells prior to addition of antibodies.

This sorting strategy requires a total of 5 fluorophores - one to bind the DNA, three for the different spermatogonial stem cell markers and one to distinguish cells that were alive from those that were dead before the commencement of the staining protocol. To minimize overlap between the emission spectra of the fluorophores, I used Ther-

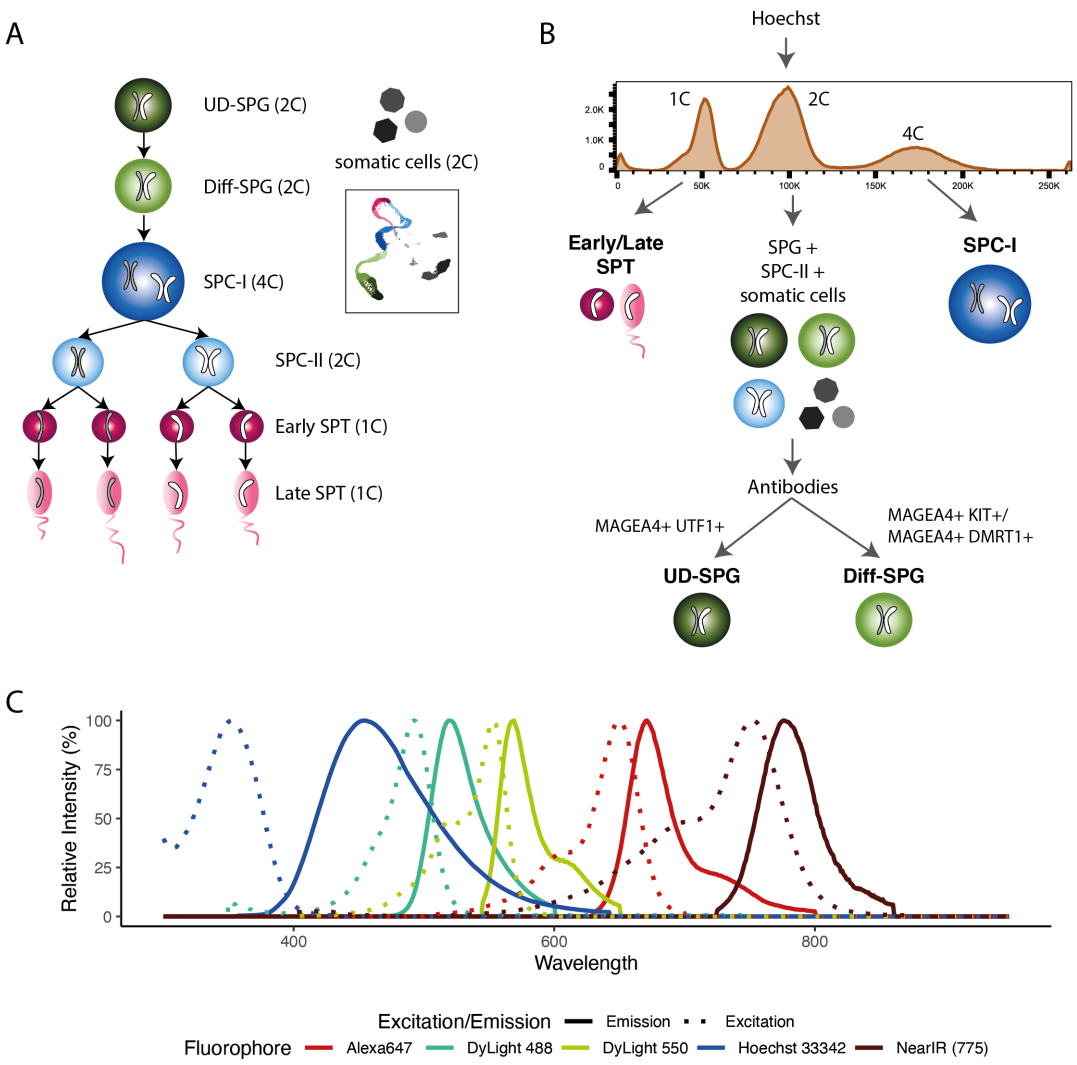



Figure 6.1: ***FACS strategy to enrich for major germ cell types in the adult human testis.*** **(A)** Schematic of major germ cell types produced during human spermatogenesis. "C" indicates the number of chromatids present within each of the cell types and the colour of the cells follows the same scheme as used in the UMAP obtained from multiome analysis of the human testis (inset). Undifferentiated spermatogonia (2C UD-SPG) commit to differentiation (2C Diff-SPG) and duplicate their DNA producing primary spermatocytes (4C SPC-I). Primary spermatocytes undergo meiosis to produce secondary spermatocytes (2C SPC-II) and subsequently haploid early and late spermatids (1C Early/Late SPT). **(B)** Schematic of the FACS strategy to utilize cellular chromatid content and cell type-specific antibodies to enrich for key germ cell types including undifferentiated spermatogonia (UD-SPG; 2C MAGEA4+ UTF1+ KIT-/DMRT1–), differentiating spermatogonia (Diff-SPG; 2C MAGEA4+ UTF1– KIT+/DMRT1+), primary spermatocytes (SPC-I; 4C) and spermatids (SPT; 1C). **(C)** The excitation and emission intensities of the fluorophores Alexa647, DyLight 488, DyLight 550, Hoechst 33342 and NearIR (775) included in the sorting strategy showing minimal overlap between their emission spectra.

moFisher's SpectraViewer tool (*https://www.thermofisher.com/order/fluorescence-spectraviewer/* n.d.) to select the following fluorophores for our design: Hoechst 33342 to stain the DNA, Dylight 488, Dylight 550 and Alexa647 to tag the antibodies and NearIR (775) to mark the dead cells (Figure 6.1C).

Since most of the working antibodies for the markers used in the sorting strategy

were of murine origin, a key prerequisite was the employment of primary antibodies directly conjugated to fluorophores. For many of the required markers, fluorophore-conjugated primary antibodies were not commercially available. In such cases, I conjugated the different fluorophores to the primary antibodies using the ThermoScientific Dylight Microscale Antibody Labelling Kits (53025, Thermo Scientific$^{TM}$) (see Methods). If the antibody was present in a different buffer or concentration, I first purified or concentrated it using a Vivaspin 500 column with 10,000 MWCO (VS0102, Sartorius) and resuspended it in sterile PBS at 1 mg/ml concentration. The antibody dilutions are included in the Appendix.

To validate the specificity of the commercial antibodies and, in the case of home-made primary conjugates, to ensure that the specificity of the antibody remained unaffected after the conjugation process, we validated the antibodies in whole mount (WM) staining of fixed human seminiferous tubules and in flow cytometry.

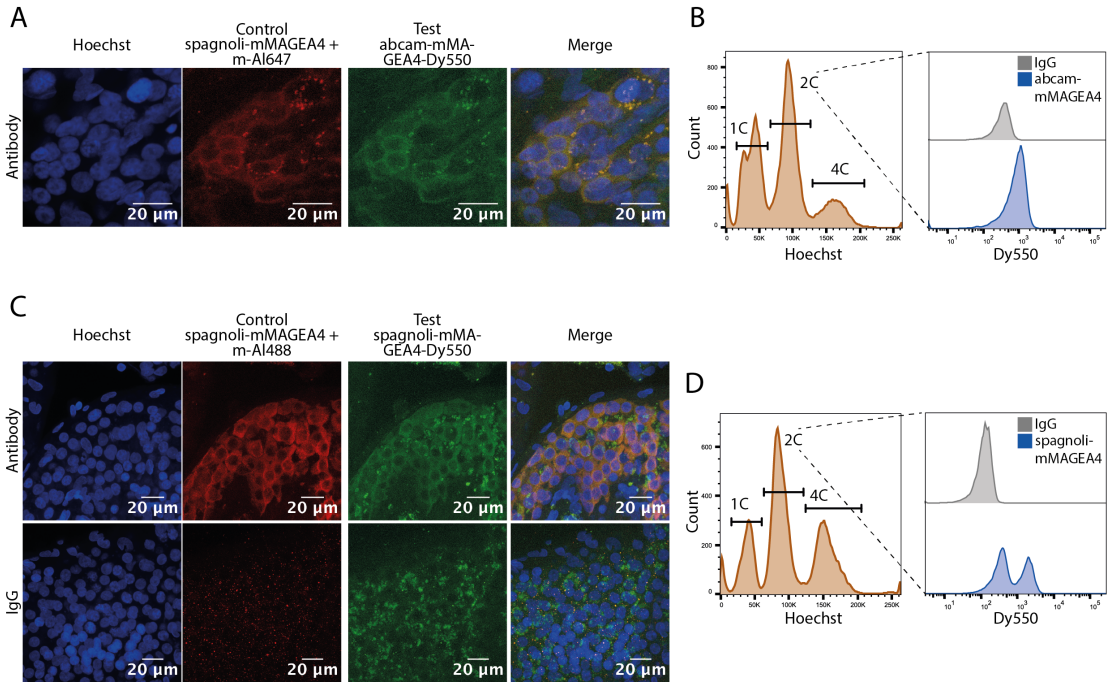



Figure 6.2: ***Testing abcam-mMAGEA4-Dy550 antibody for pan-spermatogonia cytoplasmic marker MAGEA4.*** **(A)** Whole mount (WM) staining of fixed human testicular tubules stained with Hoechst, control unconjugated spagnoli-mMAGEA4 antibody and test fluorophore-conjugated abcam-mMAGEA4-Dy550 antibody. Images were taken with Leica SP8 confocal microscope at 63x magnification and oil immersion (Images for IgG are missing). **(B)** FACS plot of human testicular single cell suspension stained with Hoechst and the test antibody abcam-m-MAGEA4-Dy550 or the matching IgG control. Cells are gated on 2C Hoechst peak before visualizing antibody staining. **(C)** and **(D)** are WM staining and FACS plot for spagnoli-mMAGEA4-Dy550 respectively.

For the pan-spermatogonia cytoplasmic marker MAGEA4, we tested two antibodies - abcam-mMAGEA4, a commercially available murine monoclonal antibody from Abcam (ab139297; Abcam) and spagnoli-mMAGEA4, a murine monoclonal antibody received as a gift from the lab of Prof. G. C. Spagnoli in National Research Council (CNR), Rome, Italy by our collaborators at CeRA (Landry et al., 2000). I conjugated both these antibodies, and their matching IgG controls, to Dylight 550. In WM staining, an unconjugated version of spagnoli-mMAGEA4 (which has earlier been proven to work in our collaborators' lab at CeRA) was used as control with an anti-mouse

secondary antibody. While abcam-mMAGEA4-Dy550 binds the same cells as control in WM staining (Figure 6.2A), it does not show a positive population compared to the IgG control within 2C cells in FACS (Figure 6.2B). However, spagnoli-mMAGEA4-Dy550 binds to the same cells as control antibody in WM staining (Figure 6.2C) and shows a positive population compared to IgG control in FACS (Figure 6.2D). Hence, we selected spagnoli-mMAGEA4-Dy550 and its matched IgG control for our final sorting.

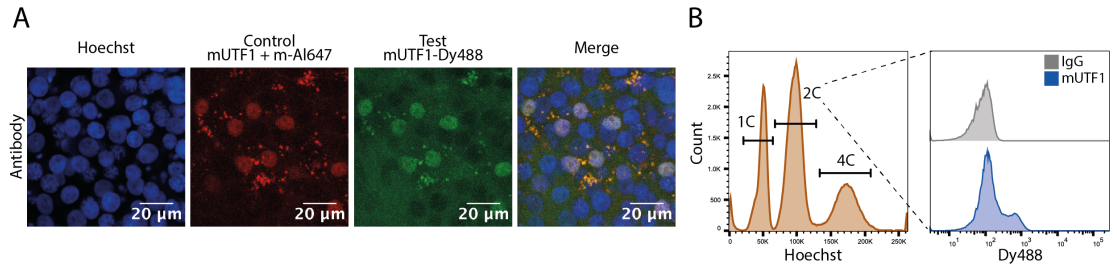



Figure 6.3: ***Testing mUTF1-Dy488 antibody for undifferentiated spermatogonia nuclear marker UTF1.*** **(A)** Whole mount staining of fixed human testicular tubules stained with Hoechst, control unconjugated mUTF1 antibody and test fluorophore-conjugated mUTF1-Dy488 antibody. Images were taken with Leica SP8 confocal microscope at 63x magnification and oil immersion (Images for IgG are missing). **(B)** FACS plot of human testicular single cell suspension stained with Hoechst and the test antibody mUTF1-Dy488 or the matching IgG control. Cells are gated on 2C Hoechst peak before visualizing antibody staining.

For undifferentiated spermatogonia nuclear marker UTF1, I conjugated the murine monoclonal UTF1 antibody from Sigma (MAB4337; Sigma), an antibody that was already shown to work in the lab of our collaborators at CeRA, and its IgG control to Dylight488. In WM staining, an unconjugated version of the primary antibody with anti-mouse secondary antibody was used as control. mUTF1-Dy488 binds the same cells as the unconjugated control (Figure 6.3A) and produces a positive population within 2C cells in FACS (Figure 6.3B). Thus, I used this to enrich for undifferentiated spermatogonia in our final sorting.

To enrich for differentiating spermatogonia, I tried the following commercially available conjugates: BL-mKIT-Al647, a murine monoclonal antibody against the surface marker KIT from Biolegend (313235; Biolegend); BD-mKIT-APC, a murine monoclonal antibody from BD Pharmingen (561118BD; BD Pharmingen$^{TM}$); and mDMRT1-Al647, a murine monoclonal antibody against nuclear marker DMRT1 from Santa Cruz (sc-377167 AF647; Santa Cruz). I also conjugated Dylight 650 to an antibody shown to work in the lab of our collaborators at CeRA, polyclonal goat KIT from R&D systems (AF332; R&D systems) to produce gKIT-Dy650 and tested the same. In WM staining, the unconjugated version of gKIT with anti-goat secondary antibody was used as control.

BL-mKIT-Al647 and BD-mKIT-APC do not show staining of a positive population in both WM staining and FACS (Figure 6.4). This could be because KIT, a surface membrane marker, gets destroyed during the fixation and permeabilization of the adult human testicular cells before antibody staining. Hence, I additionally tested BD-mKIT-APC staining in non-fixed cells (Figure 6.4D, bottom panel). Unfortunately, this did not produce a positive population either. Furthermore, Hoechst staining did not produce clear 1C, 2C and 4C peaks in non-fixed cells.

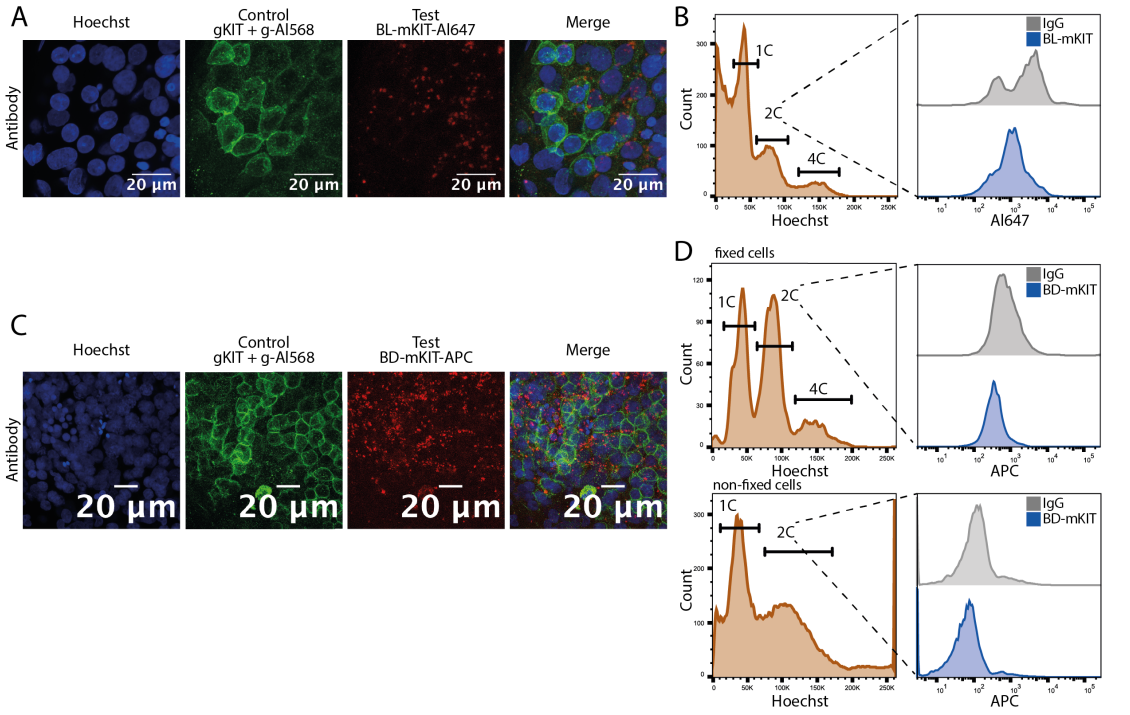



Figure 6.4: ***Testing BL-mKIT-Al647 and BD-mKIT-APC antibody for differentiating spermatogonia surface marker KIT.*** **(A)** Whole mount staining of fixed human testicular tubules stained with Hoechst, control unconjugated gKIT antibody and test fluorophore-conjugated BL-mKIT-Al647 antibody. Images were taken with Leica SP8 confocal microscope at 63x magnification and oil immersion (Images for IgG are missing). **(B)** FACS plot of human testicular single cell suspension stained with Hoechst and the test antibody BL-mKIT-Al647 or the matching IgG control. Cells are gated on 2C Hoechst peak before visualising antibody staining. **(C)** and **(D)** are WM staining and FACS plot (top panel- fixed cells; bottom panel - non-fixed cells) for BD-mKIT-APC respectively.

While gKIT-Dy650 bound to the same cells as the control unconjugated gKIT in WM staining (Figure 6.5A), it did not produce a positive population in both fixed and non-fixed cells in FACS (Figure 6.5B). Finally, mDMRT1-Al647 bound to the same cells as control antibody gKIT in WM staining (Figure 6.5C) and produced a positive population in 2C cells in FACS (Figure 6.5D). Hence, mDMRT1-Al647 was used to enrich for differentiating spermatogonia population in the final sorting.

To ensure that the antibodies do not cross-react when used together, we co-stained a fixed human seminiferous tubule with spagnoli-mMAGEA4-Dy550, mUTF1-Dy488 and mDMRT1-Al647 (Figure 6.6A). As expected, mUTF1-Dy488 and mDMRT1-Al647 bound to mutually exclusive populations of spagnoli-mMAGEA4-Dy550 positive cells.

Figure 6.6B and C depict the final gating strategy used to obtain the cells of interest. Human testicular cells were stained with LIVE/DEAD$^{TM}$ Fixable Dead Cell Stain Kit Ex: 633/635nm; Em:775nm (L34961; Invitrogen$^{TM}$). Then, they were fixed and permeabilized with FIX & PERM$^{TM}$ Cell Permeabilization Kit (GAS003;Invitrogen$^{TM}$). Subsequently, the cells were stained with the fluorophore-conjugated intracellular primary antibodies of spagnoli-mMAGEA4-Dy550, mUTF1-Dy488, mDMRT1-Al647 and the DNA stain Hoechst 33342 NucBlue$^{TM}$ Live ReadyProbes$^{TM}$ Reagent (R37605; Invitrogen$^{TM}$). Figure 6.6B depicts the gating sequence to capture live cells from the stained testicular cells. Cells were first selected from background debris (Gating: Cells). Then, single cells were selected from multiplets (Gating: Single Cells-1 and

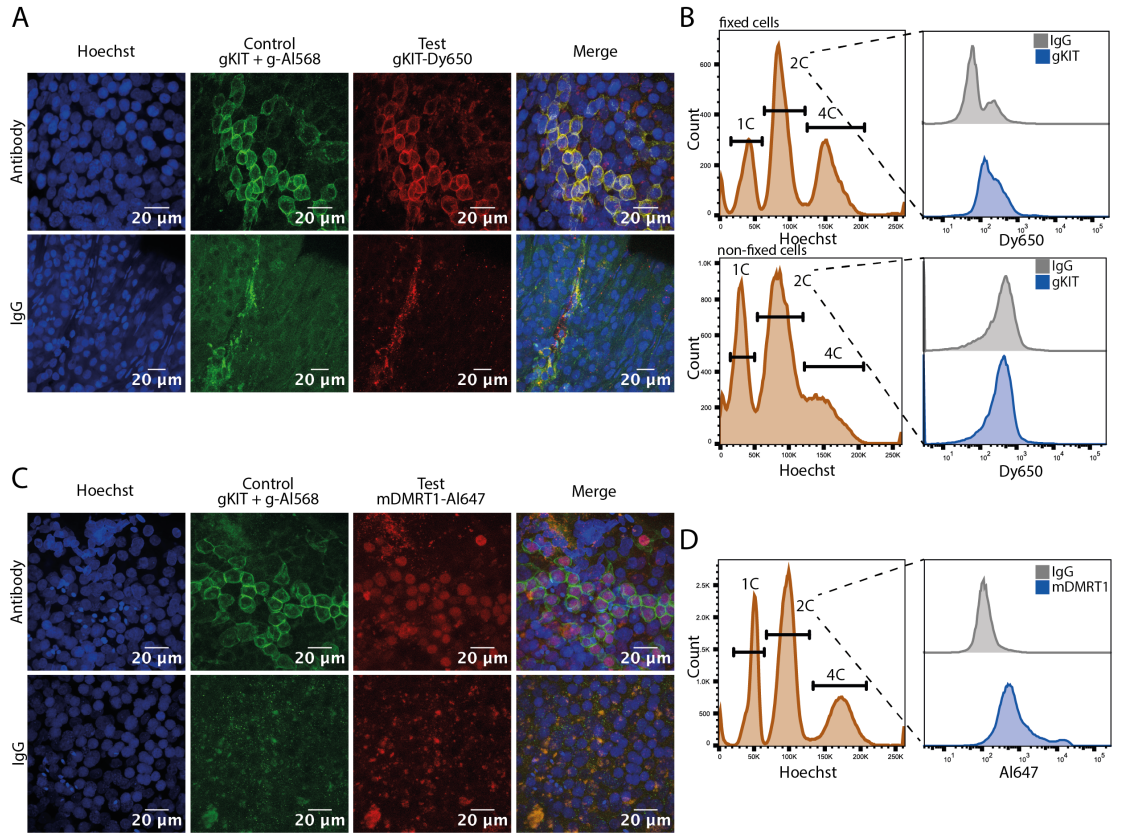



Figure 6.5: ***Testing gKIT-Dy550 and mDMRT1-Al647 antibody for differentiating spermatogonia surface marker KIT and nuclear marker DMRT1.*** **(A)** Whole mount staining of fixed human testicular tubules stained with Hoechst, control unconjugated gKIT antibody and test fluorophore-conjugated gKIT-Dy650 antibody. Images were taken with Leica SP8 confocal microscope at 63x magnification and oil immersion. **(B)** FACS plot of human testicular single cell suspension stained with Hoechst and the test antibody gKIT-Dy650 or the matching IgG control (top panel- fixed cells; bottom panel - non-fixed cells). Cells are gated on 2C Hoechst peak before visualizing antibody staining. **(C)** and **(D)** are WM staining and FACS plot for mDMRT1-Al647 respectively.

-2). Based on the signal of live/dead stain, cells that were alive at the time of fixation were selected (Gating: Live Cells). Figure 6.6C depicts the gating used on these live cells after staining with antibodies (top panel) or IgG controls (bottom panel). Spermatids and primary spermatocytes are selected based on Hoechst staining (Gating: 1C and 4C respectively). Within 2C cells, SPG are selected (Gating: MAGEA4+) and within the SPG population, undifferentiated spermatogonia (UD-SPG) (Gating: UTF1+ DMRT1-) and differentiating spermatogonia (Diff-SPG) (UTF1- DMRT1+) populations are selected. Gating for UTF1 and DMRT1 are drawn based on all 2C cells in IgG control (Figure 6.6C, last panel in the bottom).

We applied this sorting strategy to human testicular cells obtained from three adults exhibiting healthy sperm development. Figure 6.7 summarizes the number of cells I obtained from each biological replicate used in this study. In general, I obtain $\sim$ 1.5% undifferentiated spermatogonia, $\sim$ 1% differentiating spermatogonia, $\sim$ 8% primary spermatocytes and $\sim$ 10-15% spermatids from a starting material of 500,000 - 6,000,000 human testicular cells.

For Replicate 1, I additionally included 100,000 4C and 200,000 1C cells from a fourth sorting experiment for the Micro-C library generation.

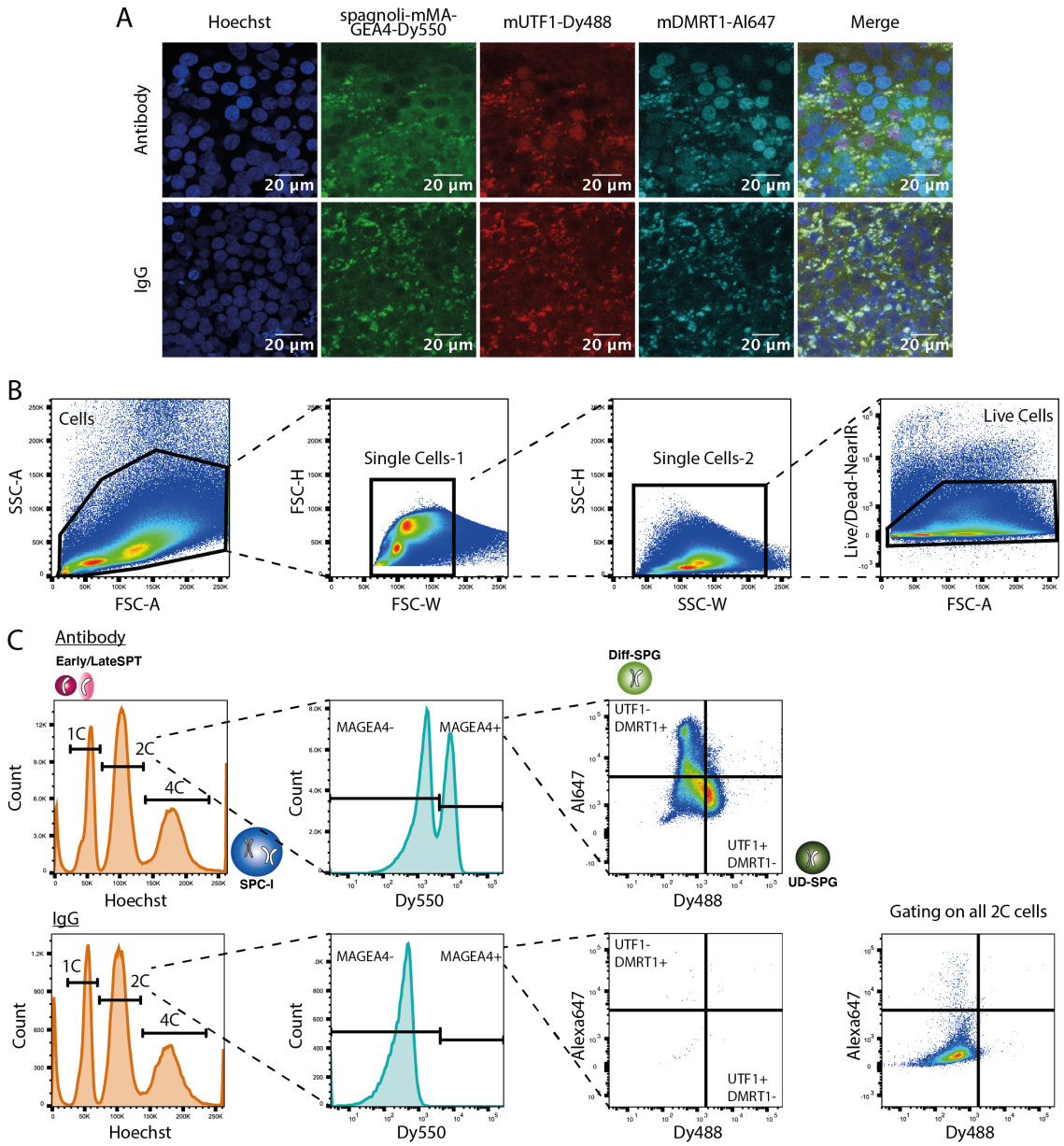



Figure 6.6: ***FACS gating sequence to enrich for key germ cells from the adult human testis.*** **(A)** Whole mount staining of fixed human testicular tubules co-stained with Hoechst, spagnoli-mMAGEA4-Dy550, mUTF1-Dy488, mDMRT1-Al647 and their IgG control. Images were taken with Leica SP8 confocal microscope at 63x magnification and oil immersion. **(B)** FACS gating sequence to capture live cells. **(C)** FACS gating sequence to capture 1C spermatids (SPT), 4C primary spermatocytes (SPC-I) and 2C MAGEA4+ UTF1+ DMRT1- undifferentiated spermatogonia (UD-SPG) and 2C MAGEA4+ UTF1- DMRT1+ differentiating spermatogonia (Diff-SPG) (top panel: antibody staining; bottom panel: IgG staining).

**Contributions** Establishment of the FACS strategy was a joint venture with Sara Di Persio (CeRA). As part of this, I designed the overall FACS strategy. For validation of antibodies, I optimized and performed the conjugation of antibodies to fluorophores (Lara Marie Siebert-Kuss at CeRA continued using this protocol to conjugate new batches of antibodies for sorting the same populations of cells for her project investigating DNA methylation changes across adult human spermatogenesis). Sara Di Persio prepared the WM staining and Sara and I captured the microscopy images together. Finally, I processed the data in Fiji ImageJ. Sara Di Persio and I also op-

| Normal Rep 1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Cell Type | FACS Population | FACS1 | | FACS2 | | FACS3 | | FACS4 | | Total |
| | | No. of Cells | % of sub-population | No. of Cells | % of sub-population | No. of Cells | % of sub-population | No. of Cells | % of sub-population | |
| | Input Cells | 1,000,000 | | 500,000 | | 1,000,000 | | - | | |
| UD-SPG | 2C MAGEA4+ UTF1+ DMRT1- | 17,420 | 1.7 | 6,833 | 1.4 | 13,713 | 1.4 | - | - | 37,966 |
| Diff-SPG | 2C MAGEA4+ UTF1- DMRT1+ | 10,832 | 1.1 | 7,187 | 1.4 | 10,958 | 1.1 | - | - | 28,977 |
| SPC-I | 4C | 85,703 | 8.6 | 49,960 | 10.0 | 123,252 | 12.3 | 100,000 | - | 358,915 |
| SPT | 1C | 117,660 | 11.8 | 92,938 | 18.6 | 218,889 | 21.9 | 200,000 | - | 629,487 |

| Normal Rep 2 | | | | |
|---|---|---|---|---|
| Cell Type | FACS Population | FACS1 | | Total |
| | | No. of Cells | % of sub-population | |
| | Input Cells | 6,000,000 | | |
| UD-SPG | 2C MAGEA4+ UTF1+ DMRT1- | 79,837 | 1.3 | 79,837 |
| Diff-SPG | 2C MAGEA4+ UTF1- DMRT1+ | 81,663 | 1.4 | 81,663 |
| SPC-I | 4C | 431,517 | 7.2 | 431,517 |
| SPT | 1C | 523,339 | 8.7 | 523,339 |

| Normal Rep 3 | | | | | | |
|---|---|---|---|---|---|---|
| Cell Type | FACS Population | FACS1 | | FACS2 | | Total |
| | | No. of Cells | % of sub-population | No. of Cells | % of sub-population | |
| | Input Cells | 500,000 | | 3,000,000 | | |
| UD-SPG | 2C MAGEA4+ UTF1+ DMRT1- | 3,868 | 0.8 | 39,992 | 1.3 | 43,860 |
| Diff-SPG | 2C MAGEA4+ UTF1- DMRT1+ | 1,943 | 0.4 | 20,924 | 0.7 | 22,867 |
| SPC-I | 4C | 20,227 | 4.0 | 188,256 | 6.3 | 208,483 |
| SPT | 1C | 5,205 | 1.0 | 69,035 | 2.3 | 74,240 |

Figure 6.7: ***Cell counts of FACS-sorted subpopulations of key germ cells from the adult human testis.*** Number of cells used as input for FACS sorting and of the subpopulations obtained for 2C MAGEA4+ UTF1+ DMRT1- (undifferentiated spermatogonia; UD-SPG), 2C MAGEA4+ UTF1- DMRT1+ (differentiating spermatogonia; Diff-SPG), 4C (primary spermatocytes; SPC-I) and 1C (spermatids; SPT) from three biological replicates. Replicates 1 and 3 were divided into and sorted 4 and 2 times respectively. 100,000 4C cells and 200,000 1C cells from a 4th sorting experiment of Replicate 1 were combined with cells obtained from FACS1, 2 and 3 for Micro-C processing.

timized the FACS protocol by testing different fixation conditions, different antibodies and their dilutions. For the samples used in this chapter, Lara Marie Siebert-Kuss and I, along with the head of FACS Facility at MPI Münster Martin Stehling, performed the FACS staining and sorting. I would like to extend my appreciation to Rebekka Stegmeyer (MPI Münster) for discussions about antibody purification and conjugation and to Malte Stasch for his guidance in processing the microscopy images.

## 6.3 Generation of Micro-C maps of 3D genome changes across adult human spermatogenesis

Micro-C is a technique that captures all chromatin interactions in the 3D space. Instead of the restriction enzyme used in the conventional *in situ* Hi-C protocol, Micro-C uses the MNase enzyme to cut the DNA into smaller fragments. This produces DNA fragments at the nucleosomal level and generates 3D genome maps at a much higher resolution compared to *in situ* Hi-C (Krietenstein et al., 2020). Furthermore, Micro-C uses a second long-range crosslinker such as EGS or DSG in addition to the formaldehyde used in *in situ* Hi-C. This captures a higher number of long-distance interactions and substantially improves the signal-to-noise ratio (Hsieh, Fudenberg, et al., 2016).

To study the 3D genome changes across adult human spermatogenesis, I generated low input Micro-C libraries from the sorted populations of human testicular cells for undifferentiated spermatogonia, differentiating spermatogonia, primary spermatocytes and spermatids. I produced libraries from three biological replicates, with the cell numbers mentioned in Figure 6.7. Prior to library generation, I first optimized the

amount of MNase enzyme required for processing each of the sorted populations on 1C, 2C and 4C cells obtained from mouse and monkey testicular suspensions (data not shown).

**Quality control** I first performed a shallow sequencing of the libraries at 1 million read pairs per library. Figure 6.8 shows the quality control analysis results of the shallow-sequenced libraries.

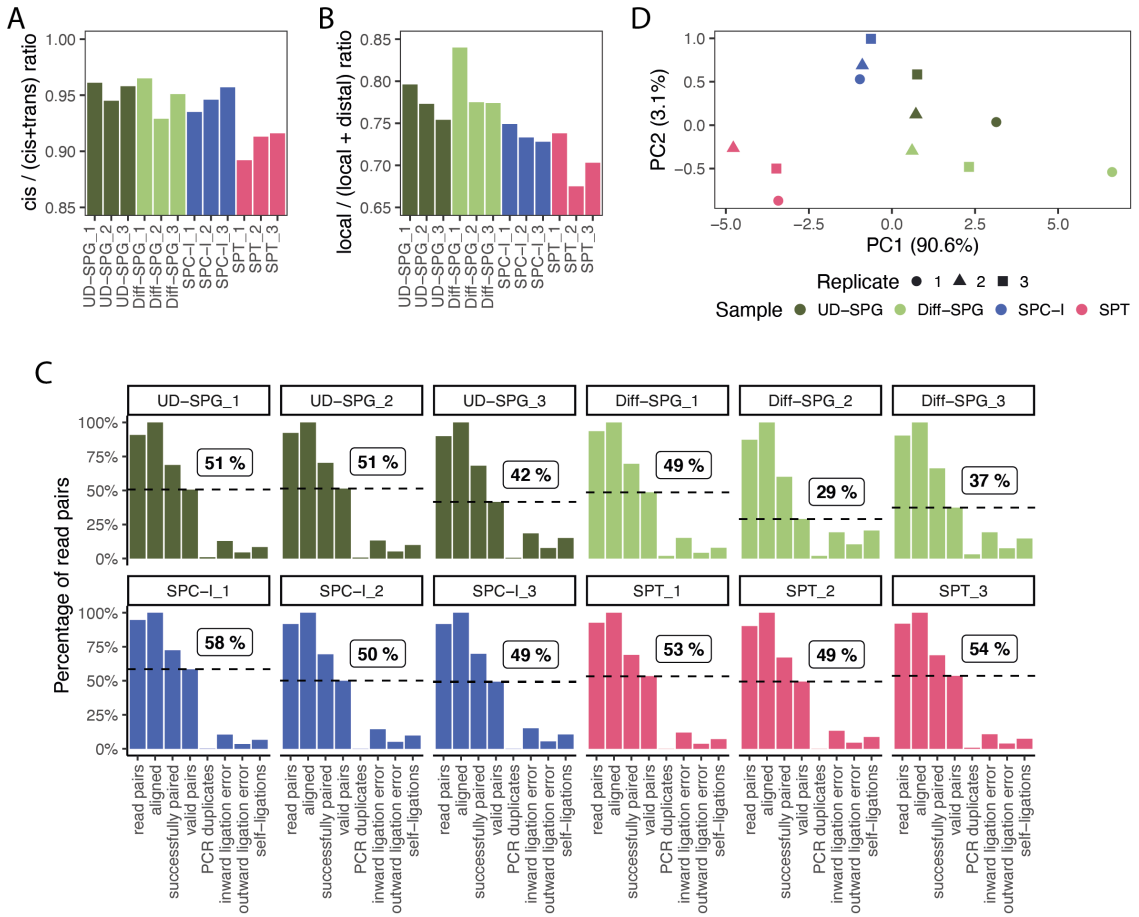



Figure 6.8: ***Quality control of Micro-C libraries across adult human spermatogenesis.*** **(A)** Ratio of intra-chromosomal contacts for Micro-C libraries of adult human undifferentiated spermatogonia (UD-SPG), differentiating spermatogonia (Diff-SPG), primary spermatocytes (SPC-I) and spermatids (SPT) from three patients exhibiting normal spermatogenesis. **(B)** Ratio of local contacts with local-distal threshold at 20 Kb for the libraries. **(C)** The percentage of reads obtained at different stages of Micro-C matrix generation for each of the libraries. The percentage of valid pairs retained for matrix generation is highlighted. **(D)** PCA plot of Micro-C matrices at 2 Mb resolution.

Cis-trans ratio and local-distal ratio are two Hi-C/Micro-C metrics that quantify the compactness of local chromatin packaging (Heinz et al., 2018). These measures are largely constant across cell types and species and hence employed as a useful proxy for evaluating the quality of the data (Lajoie et al., 2015).

Cis-trans ratio is the ratio of pairs of interacting loci which reside on the same chromosome (cis) vs those that reside on different chromosomes (trans). Since chromosomes are physically separated into chromosomal territories within the nucleus, a high-quality 3D genome map is expected to have a higher number of cis contacts

than trans contacts. Additionally, trans contacts are a mixture of true chromatin interactions and interactions that are from random ligations (Lajoie et al., 2015; Schmitt et al., 2016). Hence, a noisy 3D genome matrix is expected to return relatively fewer cis contacts. Figure 6.8A shows that all the Micro-C libraries are of high quality with $\sim$ 0.90 - 0.95 ratio of intra-chromosomal contacts (cis) to total contacts (cis + trans). The Micro-C protocol, in addition to the formaldehyde used to cross-link the DNA in the conventional Hi-C protocol, includes a second cross-linker such as disuccinimidyl glutarate (DSG) or ethylene glycol bis(succinimidylsuccinate) (EGS). The additional cross-linker has been shown to decrease the percentage of trans interactions (Akgol Oksuz et al., 2021) and thus result in a higher cis-trans ratio. Of particular interest are the spermatids libraries, which consistently show a slightly lower ratio of $\sim$ 0.90 compared to the other cell types. Since the histones are being replaced by protamines in this cell population, the slightly lower ratio could indicate more inter-chromosomal contacts in the DNA packaged around the protamines.

Along the same lines, compacted DNA is expected to have a higher number of interactions with other regions in its local vicinity (local) compared to regions that are farther away (distal). Here, we calculate the local compaction level of a region as the ratio of its local contacts (< 20 kb distance) to the sum of local and distal (> 20 kb distance) contacts. Figure 6.8B shows a local compaction ratio of $\sim$ 0.7 or higher for all the libraries. As observed in cis-trans ratio, local compaction is also reduced in spermatids, indicating a different organization of the local chromatin structure around protamines compared to histones in the other cell types.

Figure 6.8C represents the number of reads associated with each step of Micro-C library processing. The plot includes the total number of read pairs sequenced, the number of read pairs aligned to the genome and the number of read pairs successfully paired to represent the chimeric fragments generated in the Micro-C experiment. The plot also includes read pairs that are filtered out due to being PCR duplicates or representing ligation errors/self-ligation. Ligation errors occur when the two read pairs arise from the same fragment, instead of from two different fragments of a chimeric pair, and hence are not informative about DNA-DNA contacts from different parts of the genome (Belaghzal et al., 2017). Finally, the number and percentage of valid pairs that remain after all the filtration steps is shown. We retain $\sim$ 50% valid read pairs for all samples except Diff-SPG_2 (29%) and Diff-SPG_3 (37%). In line with this observation, Diff-SPG_2 was slightly over-digested after MNase treatment during library preparation (data not shown) and Diff-SPG_3 had the least number of cells for library generation (Figure 6.7). Hence, the two samples resulted in fewer valid pairs compared to the other samples.

Lastly, PCA analysis of all the libraries shows that, in general, biological replicates cluster together (Figure 6.8D). undifferentiated spermatogonia and differentiating spermatogonia clusters are found close to each other. Next to them lies the primary spermatocytes cluster. The spermatids replicates are the farthest away from the other libraries, again, indicating possible differences in protamine-associated DNA packaging compared to the other cell types. The arrangement of the samples in PCA recapitulates the developmental progression of spermatogenesis.

In conclusion, we have generated high quality Micro-C libraries from three patients with normal spermatogenesis for the stages of undifferentiated spermatogonia, differ-

entiating spermatogonia, primary spermatocytes and spermatids across adult human spermatogenesis.

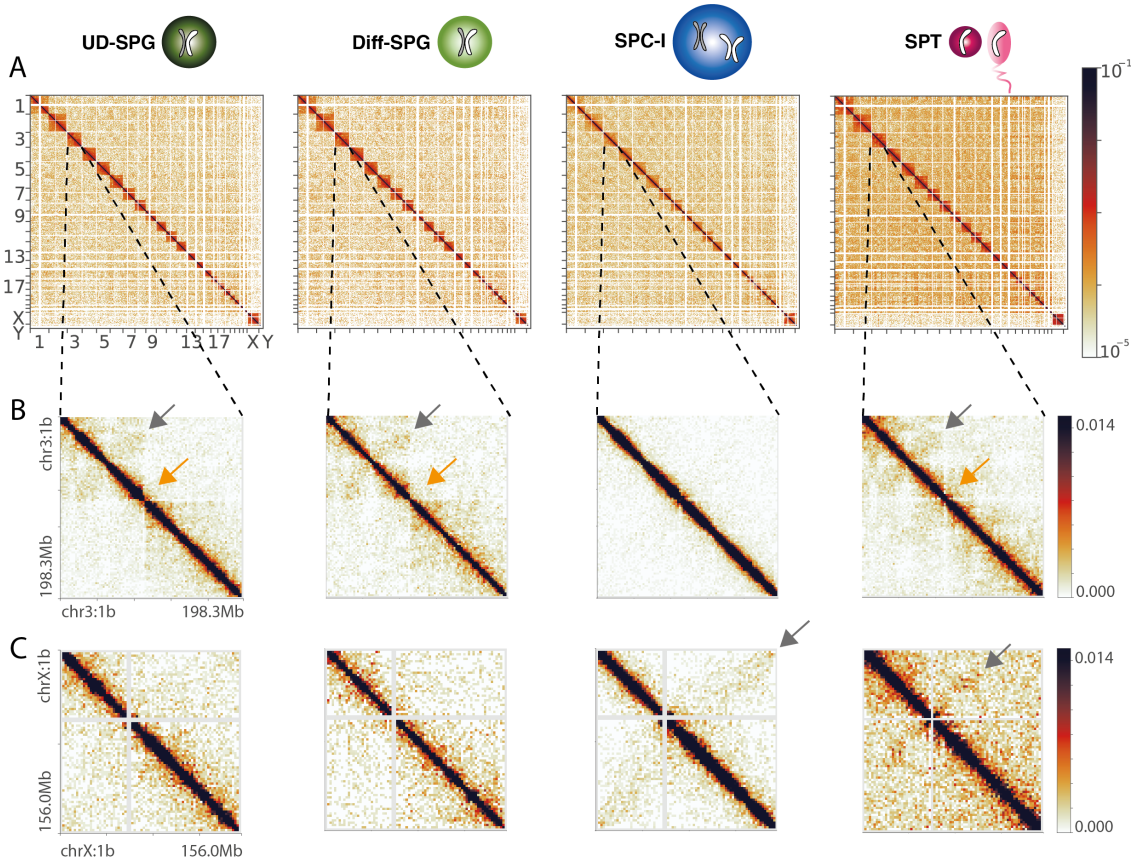



Figure 6.9: ***Micro-C contact frequency heatmaps across adult human spermatogenesis.*** **(A)** Whole genome contact frequency heatmaps of undifferentiated spermatogonia (UD-SPG), differentiating spermatogonia (Diff-SPG), primary spermatocytes (SPC-I) and spermatids (SPT) at 2 Mb resolution. Three biological replicates were merged together per stage. **(B)** Whole chromosome contact frequency heatmaps for chromosome 3 and **(C)** chromosome X at 2Mb resolution. Off-diagonal signals of interest and domain boundaries are highlighted with grey and yellow arrows respectively.

For further analysis, I merged the Micro-C datasets of the three biological replicates for each spermatogenic stage and generated contact frequency heatmaps at 2 Mb resolution. Whole genome contact frequency maps depicted in Figure 6.9A reveal the clear presence of chromosomal territories across all the stages, as seen from the squares of higher interaction frequency associated with each chromosome.

When zooming into the genome, whole chromosome contact frequency heatmap of chromosome 3 (Figure 6.9B) presents a mega domain boundary in undifferentiated spermatogonia and differentiating spermatogonia which completely disappears in primary spermatocytes and may be partially restored in spermatids (yellow arrows). All stages, except primary spermatocytes, further show a short off-diagonal signal in the first domain (grey arrows) that follows the same trend.

While meiotic primary spermatocytes still retain chromosomal territories, in general, they show very little signal away from the diagonal when zooming in at the level of the individual chromosomes. This is in line with previous reports of attenuated compartments and TADs in pachytene spermatocytes in other organisms (Patel et al., 2019; Wang, Wang, et al., 2019; Vara et al., 2019; Alavattam et al., 2019; Luo et al., 2020).

Furthermore, the loss of structure in meiotic primary spermatocytes also resembles the mitotic 3D genome organization (Gibcus et al., 2018; Naumova et al., 2013).

Very interestingly, the X chromosome (Figure 6.9C) has a strong signal perpendicular to the diagonal (grey arrow) only in primary spermatocytes. It also shows a similar signal in spermatid (grey arrow) that is slightly offset from the signal in primary spermatocytes. While the origin of the signal in spermatid is unclear, the signal in primary spermatocytes could represent the folding of the sex chromosomes to form the XY body during meiosis. Wang, Wang, et al., 2019 also observe specific off-diagonal chromatin interactions in Hi-C datasets of chromosome X in pachytene spermatocytes in mouse and monkey. While the Hi-C interactions are much fainter, the authors also attribute the signal to the folding of X chromosome in the XY body. Further analysis of the Micro-C libraries at a higher resolution is required to confirm this possibility.

The resolution of the preliminary dataset restricts further observation of loops and domain boundaries in the Micro-C libraries. Since the shallow-sequenced datasets confirm high quality of the libraries and successfully capture 3D genome changes across spermatogenesis, we have now performed deep-sequencing at 2 billion read pairs/library. The deep-sequenced libraries are currently being mapped to the genome. Since the datasets are considerably larger, it requires more processing time and thus the results have not been included in the thesis at this stage.

In summary, we have generated the first Micro-C maps across adult human spermatogenesis and observe changes in 3D genome organization across both autosomes and sex chromosomes.

**Contributions** The low input Micro-C protocol is a novel and unpublished adaptation of the method used in Akgol Oksuz et al., 2021 and Hsieh, Cattoglio, et al., 2020 by the Vaquerizas group. The protocol was optimized by lab members Noura Maziak (MRC LMS) and Dr. Virginie Tissieres (MRC LMS; former member) to work with as few as 10,000 cells of Drosophila embryos and mESCs respectively. I would like to extend my appreciation to them for their guidance during my implementation of the same in human testicular samples.

## 6.4 CTCF footprint decreases during adult human spermatogenesis

The lack of TADs in human mature sperm has been ascribed to an absence of CTCF at the protein level. However, it still shows the presence of RAD21 protein.

In line with this observation, our snATAC-seq + snRNA-seq multiome dataset of the adult human testis also shows a reduction in CTCF footprints across germ cell differentiation, reaching minimal values from Cluster 8: Pachytene/Diplotene to Cluster 10: Late spermatids (Figure 6.10).

Thus, further exploration of the spermatid architecture is required to confirm the recovery of domain structures in post-meiotic spermatogenesis, and if so, the possible mechanisms supporting the same.

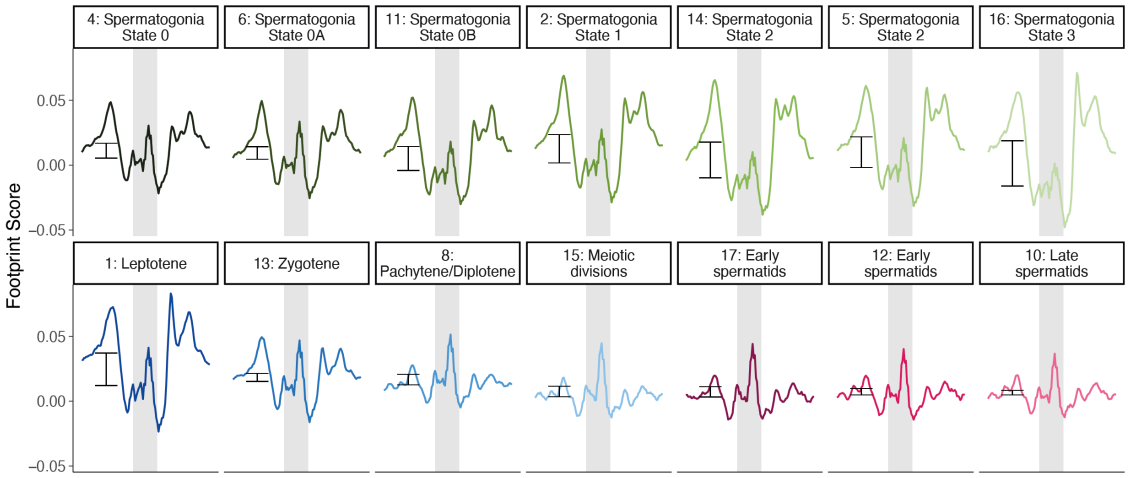



Figure 6.10: ***CTCF/CTCFL footprint in the adult human testicular germ cells.*** Grey-shaded area represents the centre of the motif and The error bars depict the footprint depth or the difference in accessibility between centre and flank of the motif binding region. Footprint depth is highest in Cluster 16: Spermatogonia State 3 and Cluster 1: Leptotene and is almost negligible from Cluster 8: Pachytene/Diplotene to Cluster 10: Late Spermatids.

## 6.5 Discussion

In this set of experiments, we have, for the first time, isolated key developmental stages of adult human spermatogenesis including undifferentiated spermatogonia, differentiating spermatogonia, meiotic primary spermatocytes and haploid spermatids. To do this, we have established a FACS-based strategy. Importantly, we have validated the antibodies used in the sorting protocol in whole mount staining.

This strategy is pivotal in studying the heterogeneous populations of the testis, which is additionally challenging in humans due to the limited number of cells obtained from biopsy tissues. Our strategy requires much lower starting material than currently prevalent methods such as STA-PUT. STA-PUT uses a linear BSA gradient through which individual cell types sediment with different sedimentation velocities based on their cell size and density. On average, we started with $\sim$100 mg tissue and 1-2 million cells per sample. However, STA-PUT requires an optimal number 700-800 million cells as starting material (Bryant et al., 2013). Reports of using STA-PUT with fewer cells are inconsistent. Y. Liu et al., 2015 state that they use $\sim$ 5 million human testicular cells as input for STA-PUT. However, they also mention starting with 6 grams of biopsy tissue and pooling 3-4 biopsies for each isolation. Moreover, unlike our strategy, the authors are unable to separate undifferentiated spermatogonia from differentiating spermatogonia using STA-PUT.

The FACS sorting strategy established in this study has been used to investigate changes in DNA methylation across the aforementioned cell types by our collaborators at CeRA. Hence, apart from Micro-C, the isolation protocol has been greatly amenable with at least one other genomic technique.

However, our strategy has some limitations. The first one is that, while we do capture undifferentiated and differentiating spermatogonia with UTF1 and DMRT1 antibodies respectively, only a fraction of these two cell types express these markers. The

remaining subpopulations of undifferentiated and differentiating spermatogonia, example GFRA1+ State 1 undifferentiated spermatogonia or KIT+ differentiating spermatogonia, will not be captured in our analysis. A possible solution is to use more antibodies targeting other subpopulations of spermatogonia. However, FACS sorting instruments have a limited number of sorting streams (for example, BD FACS Aria Fusion used in this project contains 4 sorting streams), and so, one would have to perform sequential sorting.

Another limitation of our strategy is the lack of resolution within the meiotic stages and within the spermatid stages. Most of the published studies observe maximal changes in 3D genome organization in pachytene spermatocytes when recombination occurs. The signal from the 4C population in this study would be a mix of leptotene, zygotene, pachytene and diplotene. Hence, while very strong changes will still be captured, the specific differences between the meiotic stages will be averaged out and not captured. Similarly, we will also not capture the 3D genome changes occurring across haploid cell maturation from round to elongating spermatids.

Lastly, the strong fixation and permeabilization steps employed in this sorting strategy may not be readily amenable with all protocols for downstream processing. Our attempts at using antibodies against membrane markers like KIT were not successful (Figure 6.4 and Figure 6.5). Hence, all the markers used in the sorting strategy are intracellular, making the fixation and permeabilization steps indispensable for the antibodies to enter the cell. While we confirm that the sorting strategy can be used to generate Micro-C and DNA methylation datasets across adult human spermatogenesis, it may still require more optimization to be used with, for example, CUT&RUN, a technique that is more sensitive to fixation parameters (Miura and Chen, 2020).

Using the sorted fractions of the adult human testis, we have produced the first high quality Micro-C maps across adult human spermatogenesis. The Vaquerizas lab has adapted the Micro-C protocol to be used with as few cells as 10,000 (unpublished protocol). This optimization was crucial to produce the 3D genome maps across adult human spermatogenesis since the cells of interest form a very small fraction of the testis (1-15%) (Figure 6.7). This is especially striking for differentiating spermatogonia, where I have generated Micro-C libraries from as low as $\sim$ 20,000 cells. The Micro-C libraries generated in this study exhibit characteristics of high-quality datasets including a high ratio of cis-trans contacts and local-distal contacts. I produced three biological replicates for each of the sorted populations of human testicular germ cells, and the Micro-C matrices of the three replicates cluster together in PCA analysis.

Micro-C contact frequency maps of the merged biological replicates at 2 Mb resolution present chromosomal territories in all the spermatogenic stages. In contrast, maps of the individual chromosomes capture changes across germ cell development, including a complete loss of off-diagonal signal in primary spermatocytes in autosomes. This is similar to observations made in Hi-C datasets of mouse and monkey spermatogenesis (Patel et al., 2019; Wang, Wang, et al., 2019; Vara et al., 2019; Alavattam et al., 2019; Luo et al., 2020). However, female germ cells undergoing meiosis still retain typical 3D genome structures in meiosis I, the developmental stage equivalent to that of primary spermatocytes. Oocytes only show a loss of chromatin structure during meiosis II, when secondary spermatocytes are already produced in spermato-

genesis (Tian et al., 2022). Hence, 3D genome changes in germ cell development seem to follow different temporal dynamics between males and females.

While the mitotic 3D genome also shows a reduction in off-diagonal signal similar to meiotic primary spermatocytes (Gibcus et al., 2018; Naumova et al., 2013), the interaction probability across different genomic distances differs. Specifically, the P(s) curve depicting the relationship between the interaction probability and the genomic distance shows a steep drop at $\sim$ 30 Mb for mitosis and at $\sim$ 8 Mb for meiosis (Luo et al., 2020). This difference indicates that mitosis and meiosis have distinct mechanisms of chromatin condensation. Furthermore, unlike the mitotic genome, meiotic genome is transcriptionally active from pachytene onwards (Monesi, 1964; Paronetto et al., 2011) and exhibits A/B compartmentalization (Luo et al., 2020). Using Micro-C, unlike Hi-C, we will be able to capture possible micro-TADs and gene-level folding (Hsieh, Cattoglio, et al., 2020) that contribute to the observed transcriptional activity during meiosis.

Another interesting observation in meiotic primary spermatocytes is the presence of a strong signal perpendicular to the diagonal in the X chromosome. This signal represents contacts between the two ends of the X chromosome and possibly captures the unique folding of the sex chromosomes into the XY body during MSCI. A similar pattern of contacts has also been observed in the X chromosome in mouse and monkey pachytene spermatocytes (Wang, Wang, et al., 2019).

A key motivation behind the investigation of chromatin changes in human spermatogenesis is the apparent difference in 3D genome organization between human and mouse mature sperm. The human sperm, unlike that of other model organisms, was reported to lack TADs and this absence was attributed to the lack of CTCF in the sperm (Chen, Ke, et al., 2019). In line with these observations, the snATAC-seq + snRNA-seq multiome data of the adult human testis generated in this project shows reducing CTCF footprint during human spermatogenesis. However, the loss of TADs in mouse pachytene spermatocytes occurs despite the presence of CTCF at this stage (Patel et al., 2019; Vara et al., 2019). Hence, CTCF may not be the only determinant of TAD formation in spermatogenesis. Furthermore, male germ cells express the testis-specific paralog of CTCF called CTCF-like (CTCFL) or BORIS (Loukinov et al., 2002; Sleutels et al., 2012). CTCFL does not have an impact on compartmentalization nor does it rescue TADs after CTCF depletion. However, it plays an indirect role in 3D genome organization by disrupting CTCF-mediated genome folding (Nishana et al., 2020). Hence, the Micro-C datasets, when integrated with genome-wide protein binding assays of CTCF and CTCFL, could further disentangle the mechanisms leading to the loss of TADs in mature human sperm.

Contrary to previous observations of the 3D genome organization in mouse, Vara et al., 2019 report a lack of TADs in mouse sperm. Recent studies also indicate that the presence of TADs in mouse sperm as reported in the earlier studies might be erroneous due to contamination from cell-free DNA(Yin et al., 2022). The authors propose the inclusion of DNAse I and the reducing agent DTT to remove contaminating DNA and successfully gain access to sperm DNA respectively. Our protocol to generate the 3D genome maps from human testicular cells includes both these chemicals, with DNAse I used during testicular tissue digestion and DTT used during Micro-C library preparation. Hence, our observations should provide a more accurate representation

of the 3D genome landscape during spermatogenesis.

A limitation of our current study is that the most mature stage included is that of spermatids. Spermatids undergo further differentiation to finally develop into the mature sperm stored in the epididymis of the testis. Hence, we will not capture the 3D genome changes that occur during the very final stages of sperm development in this study. However, mature sperm from semen sample is easily accessible and can be included in this study at a later stage.

After confirming the quality of the Micro-C datasets from shallow sequencing of 1 million read pairs per library, we have now performed deeper sequencing of the datasets at 2 billion read pairs per library. We are currently analyzing these datasets, but since the data is considerably larger and requires more processing time, the results have not been included in the thesis at this stage. In summary, we have generated the first high resolution 3D genome maps across human spermatogenesis that capture dynamic patterns of chromatin organization during this process, and we are currently conducting a detailed examination of the same.

# Chapter 7

# Summary and conclusions

In this thesis, I have explored the changes in the gene regulatory landscape and the resulting gene expression patterns that are characteristic of the different stages of adult human spermatogenesis.

In Chapter 3, I presented the optimization and generation of joint profiles of chromatin accessibility and gene expression from the same single cells of the adult human testis. In Chapter 4, I explored the human spermatogonial compartment and characterized the different subpopulations present. In Chapter 5, I achieved the separation of spermatids into X- or Y-chromosome-bearing and compared their genome and transcriptome. In Chapter 6, I presented the generation and preliminary characterization of the 3D chromatin organization maps during key stages of human sperm development.

In this chapter, I will provide a summary of the main results of the thesis and their implications. I will also discuss future perspectives that stem from this work and that could further dissect the gene regulatory landscape that drives human spermatogenesis.

## 7.1 Joint profiling of gene expression and chromatin accessibility in the adult human testis at the single cell level

In Chapter 3, I introduced the generation of snATAC + snRNA multiomic data from the same single cells of the adult human testis. For this, I employed the Chromium Next GEM Single Cell Multiome ATAC + Gene Expression kit from 10x Genomics which requires permeabilized nuclei as input. I optimized different parameters of the nuclei isolation protocol including lysis buffer composition, duration of lysis, centrifugation speed and pre-treatment of the testicular single cell suspension. I found that a 30-second incubation in CUT&Tag lysis buffer resulted in maximum recovery of high-quality nuclei from the testicular single cell suspension. Under these conditions, I still retained a few unlysed elongating spermatid heads in the nuclei suspension. To overcome this, I tried fixing the cells with methanol prior to nuclei isolation, to bring the heterogeneous cell population to the same level of sensitivity to lysis conditions, but

this did not produce successful snATAC-seq libraries. Thus, despite isolating nuclei from fresh tissue which leads to a depletion of nucleus from elongating spermatids, the generated multiome data captured the entire known repertoire of the adult human testis. Using known markers, we were able to identify a continuum of germ cells at different stages of development and the supporting somatic cells present in the adult human testis. Our data captures stage-specific gene expression and chromatin accessibility profiles.

As expected from the observations during nuclei isolation, our snRNA-seq captured a higher proportion of spermatogonia and lower proportion of spermatids compared to published scRNA-seq datasets (Guo, Grow, Mlcochova, et al., 2018; Di Persio, Tekath, et al., 2021). The higher proportion of spermatogonia in our dataset facilitates a deeper examination of the stem cell compartment in the adult human testis. Moreover, the gene expression patterns of all the cell clusters in the snRNA-seq data show high significant correlations with the corresponding clusters in the published scRNA-seq studies.

Hence, the high-quality joint profiles of snATAC-seq and snRNA-seq generated for the adult human testis can be used to investigate the gene regulatory landscape of healthy sperm development and form a basis to explore possible misregulation in infertility.

## 7.2 Characterization of the human spermatogonial compartment and its interaction with the niche

In Chapter 4, I performed a focussed analysis on the human spermatogonial compartment and identified eight new subclusters. This includes a cluster at the point of transition from undifferentiated spermatogonia to differentiating spermatogonia. The transcriptome of this cluster resembles an undifferentiated state, while the accessibility at transcription factor motifs resembles a more differentiated state. This cluster possibly captures a cell state that is just preparing for differentiation. I also identified a subpopulation of proliferating State 1 spermatogonia that has not been captured in previous scRNA-seq studies (Guo, Grow, Mlcochova, et al., 2018; Sohni et al., 2019) but has been observed in immunofluorescence staining of the adult human seminiferous tubule (Di Persio, Saracino, et al., 2017).

Dynamic patterns of genome-wide accessibility reveal that the number of accessible peaks decreases from undifferentiated spermatogonia until the beginning of differentiation, after which it increases again. The pattern possibly recapitulates the switch in transcriptional programs that occurs at the transition from undifferentiated to differentiating spermatogonia.

Additionally, when the cells are clustered based on the joint similarity of their gene expression and chromatin accessibility profiles, the undifferentiated spermatogonia form two clusters (which we call State 0A and State 0B) that are organized into two separate branches, each arising from the most naive population of State 0 and ending at State 1. Velocity analysis is inconclusive in inferring the relationship between the different undifferentiated spermatogonia states. However, State 0B is enriched for mitochondrial transcripts and proliferative cells compared to State 0A, and the two

clusters could possibly capture different metabolic or physiological states of the same cell type. The branches also exhibit differences in niche signalling which includes the relative enrichment of GAS6-MERTK in State 0A and PSAP-AR in 0B, suggesting the existence of the states in different microenvironments or at different time points along the cycle of a seminiferous tubule. This included ligands or receptors that were also found to be misregulated in the infertility leading to asthenozoopspermia.

In summary, this chapter contributes details pertaining to the dynamics of chromatin accessibility and niche signalling to the rapidly growing field of spermatogonia characterization in the adult human testis.

## 7.3 Genomic and transcriptomic differences between X- and Y-spermatids

In Chapter 5, I separated haploid spermatids into chromosome X- or Y-bearing using the snATAC-seq data and performed a comprehensive comparison of the genomic and transcriptomic profiles of the two. In general, genes that overlap differentially accessible peaks between X- and Y-spermatids show a higher expression in the cluster of enriched accessibility than genes with no overlap, and vice-versa. All the genes and peaks that showed enrichment in the X-spermatids were present on the X chromosome. However, the Y-spermatids showed differential gene expression also on autosomes. Early Y-spermatids show enriched expression of the DNA-binding zinc-finger protein ZNF28. *ZNF28* is located on the autosomal chromosome 19 and has a putative role in the silencing of the hominoid-specific retrotransposon family SVA (Fukuda, Makino, et al., 2022). The pattern of early Y-spermatid-specific *ZNF28* expression with motif accessibility in both late X- and Y-spermatids indicates a possible distribution of the protein from Y-spermatids to X-spermatids via the cytoplasmic bridges connecting them, to achieve the silencing of SVAs in all spermatids.

Thus, this chapter provides a basis for further exploration of possible differences in the regulation of repetitive elements between X- and Y-spermatids. Optimizing the nuclei isolation protocol to retain more spermatids in the dataset could identify more differences that could have been missed in our current analysis.

## 7.4 The dynamic 3D genome organization across adult human spermatogenesis

In Chapter 6, I presented the first sorting strategy that enriches for key germ cell types during human sperm development including undifferentiated spermatogonia, differentiating spermatogonia, meiotic cells and haploid spermatids. I discussed the process of establishment of the sorting strategy including the generation of fluorophore-conjugated antibodies, and their validation in flow cytometry and immunofluorescence staining of fixed human seminiferous tubules. Using a novel adaptation (unpublished methodology) of the Micro-C protocol to process low cell numbers, I generated high resolution 3D genome maps across these stages from three individuals displaying healthy sperm development. The libraries display characteristics of high quality in-

cluding cis-trans ratio > 0.9-0.95, local-distal ratio > 0.7 and the formation of clusters based on the developmental stage in PCA analysis. Similar to observations in other organisms (Patel et al., 2019; Wang, Wang, et al., 2019; Vara et al., 2019; Alavattam et al., 2019; Luo et al., 2020), the 3D genome maps in human meiotic spermatocytes show very low signal away from the diagonal, capturing the unique arrangement of the chromosomes during this process. Very interestingly, we observe signal perpendicular to the diagonal only in the meiotic X chromosome, which indicates the distinctive folding of the sex chromosomes into the XY body during MSCI.

Lastly, the observed lack of TADs in human mature sperm has been attributed to the absence of CTCF in them. In line with these observations, footprinting analysis from our multiome data reveals the strongest footprint for CTCF in differentiating spermatogonia State 3 and leptotene spermatocytes and almost negligible signal from pachytene/diplotene onwards.

In summary, the results from this chapter lay the groundwork for deeper understanding of the changes in the 3D genome organization that accompany or drive human spermatogenesis.

## 7.5 Future perspective

The results of the thesis can be used to draw further insights into many more interesting phenomena that occur during human spermatogenesis.

### 7.5.1 Spermatogonial homeostasis

Within the spermatogonial compartment, our analysis reveals a subpopulation whose transcriptome reflects an undifferentiated state of development while the accessibility within TF motifs reflects a differentiated state. This cluster potentially reflects a cell state that is just about to enter differentiation and one could further explore the transcription factors and their target genes that set off this cascade of differentiation.

Another interesting aspect to explore within the spermatogonial compartment is the comparison between the mouse and the human testis as they are found to show considerable differences. The mouse spermatogonia can be morphologically distinguished into many subgroups including $A_{single}$, $A_{paired}$ and $A_{aligned}$, $A_{1\text{-}4}$, Intermediate and B spermatogonia (Dettin et al., 2003; Fayomi and Orwig, 2018). However, such a method of classification has revealed far fewer subgroups in the human testis, that include $A_{dark}$, $A_{pale}$ and B. Mouse and human spermatogonia also differ in terms of their proliferative capacity and proportion within the testis (Shami et al., 2020). In both species, *GFRA1*+ cells represent undifferentiated spermatogonia. However, in humans, an additional subgroup called State 0 has been found, which represents spermatogonia at a more naive state than *GFRA1*+ cells and is characterized by the expression of *PIWIL4* (Guo, Grow, Mlcochova, et al., 2018; Sohni et al., 2019; Di Persio, Tekath, et al., 2021). The longer lifespan of humans is thought to necessitate a reserve pool of spermatogonia with the ability to maintain the reproductive capacity of the species for a longer duration. Alternatively, in the mouse testis, *Gfra1*+ cells have been reported to undergo stochastic interconversion between a renewal-biased

and a differentiation-primed state to maintain a homeostatic and persistent spermatogenesis (Nakagawa et al., 2021). While similar interconversions have been proposed through computational analyses in humans as well (Guo, Grow, Mlcochova, et al., 2018), it has not been proved experimentally.

Furthermore, recent studies of the mouse testis have reported a very small fraction of *Miwi2+* (the murine homologue of *PIWIL4*) spermatogonia that clusters together with the human *PIWIL4*+ State 0 spermatogonia in scRNA-seq analysis (Shami et al., 2020). Carrieri et al., 2017 have also found a subset of *Ngn3*+ undifferentiated spermatogonia in the adult mouse testis that express *Miwi2*. While *Ngn3*+ cells form a transit amplifying population under homeostasis, the *Miwi2*+ subset was found essential for efficient regeneration of the germ cell population after cell ablation. Thus, during germ cell development, *PIWIL4*+ population in the adult human testis has been proposed to exist earlier than *GRFA1*+ cells. *Miwi2*+ cells in the adult mouse testis has been proposed to exist later than the *Gfra1*+ population, but with the capacity to de-differentiate and produce *Gfra1*+ cells during regeneration. Taken together, it is unclear if equivalent spermatogonia states exist between the two species, and if not, what distinct mechanisms are at play to maintain the reproductive capacity of the respective species.

### 7.5.2 Escape from post-meiotic persistence of MSCI

sc/snRNA-seq analysis of the testis has captured the silencing of sex chromosomes during MSCI across multiple species (Murat et al., 2023). MSCI-initiated silencing is found to persist beyond meiosis (Namekawa et al., 2006; Turner, Mahadevaiah, Ellis, et al., 2006), however, some genes are found to escape this persistent silencing through multiple mechanisms such as formation of multi-copy genes (Mueller et al., 2008), histone modification by specific factors (Sin and Namekawa, 2013), etc. Using the multiome data, one can investigate if there exists a core set of TFs or specific chromatin environment that enable the escape from post-meiotic silencing.

### 7.5.3 Post-transcriptional regulation of spermatogenesis

Spermatogenesis requires elaborate regulation of cellular identity to produce mature sperm from spermatogonia. A growing body of evidence has shown a role for post-transcriptional modifications such as RNA $m^6A$ methylation in regulating spermatogenesis. Alterations in the readers, writers and erasers of this mark have been implicated in male infertility and spermatogenic defects (S. Liu et al., 2021). The multiome data can be utilized to understand the dynamics of the $m^6A$ modifiers and the resulting differences in gene expression of the target transcripts. While this provides a starting point to explore $m^6A$ regulation of spermatogenesis, the observations will have to complemented by direct measurement of $m^6A$ methylation levels through techniques such as crosslinking and immunoprecipitation (CLIP) (Hafner et al., 2021; Ernst et al., 2023).

### 7.5.4 Dynamics of expression of transposable elements in adult human spermatogenesis

Transposable elements are especially active during male germline development due to a more relaxed epigenetic environment. Mammalian primordial germ cells (PGCs) undergo global epigenetic reprogramming in early development, including DNA demethylation and resetting of histone marks. This is accompanied by the transcription of retrotransposons and the upregulation of PIWI-piRNA pathway genes to counteract transposon activity (Gkountela et al., 2013; Guo, Yan, et al., 2015; Tang, Dietmann, et al., 2015; Tang, Kobayashi, et al., 2016). Previous work has also shown stage-specific enrichment of different classes of transposable elements during adult human spermatogenesis (Guo, Grow, Mlcochova, et al., 2018). Thus, it would be interesting to utilize the multiome data to generate a map of transposable elements enrichment and the changes in the genomic landscape during this process.

The dynamics of transposable elements can also be investigated within the spermatogonial compartment, including the new subclusters identified in this study. As mentioned earlier, one of the defining features of State 0 spermatogonia in the adult human testis is the enrichment of *PIWIL4*. The PIWI proteins belong to the Argonaute family, which has traditionally been shown to play a role in protecting the integrity of the germ cell genome by suppressing the transposable elements. Guo, Nie, et al., 2020 have found a similar state called State f0 which arises from PGCs at 14 weeks post fertilization. Hence, *PIWIL4* expression in the adult human spermatogonia could indicate a residual necessity to silence transposable elements. *Miwi2* is expressed from E14.5 until shortly after birth (Aravin et al., 2008) and mice mutant for *Miwi2* show germ cell arrest at early prophase I with possible links and inappropriate activation of transposable elements (Carmell et al., 2007). While inducible knockout of *Miwi2* in the adult spermatogonia does not negatively impact homeostasis or regeneration of the germ cells, *Miwi2*$^{-/-}$ mice exhibit de-repression of intracisternal-A particle (IAP) family of endogenous retroviruses elements that inappropriately promoting the expression of neighbouring genes and novel transcripts (Vasiliauskaitė et al., 2018). It would be interesting to dissect the role of *PIWIL4* in the developing and the adult human testis.

Lastly, transposable elements expression analysis could also reveal novel regulation in spermatid development. Apart from the necessity of *Miwi2* for complete spermiogenesis in mouse testis, our comparison of X- and Y-spermatids identifies an enrichment of *ZNF28* in early Y-spermatids. Since ZNF28 is potentially associated with the silencing of SVAs (Fukuda, Makino, et al., 2022), this could suggest different dynamics of transposable elements silencing between X- and Y-spermatids.

### 7.5.5 3D genome organization across adult human spermatogenesis

Finally, as discussed in Chapter 6 of the thesis, the human mature sperm has been shown to lack TADs and this has been associated with an absence of CTCF (Chen, Ke, et al., 2019). While reports of TADs in mature sperm of other species is conflicted (Yin et al., 2022; Vara et al., 2019; Alavattam et al., 2019; Wang, Wang, et al., 2019), the developmental stages preceding mature sperm display dynamic 3D

genome organization, especially in pachytene spermatocytes. The work in this thesis includes a preliminary analysis of 3D genome organization across similar developmental stages in the adult human testis. Ongoing work in the lab includes a detailed examination of alterations in compartmentalization, TADs and loops at high resolution. Incorporating DNA-binding data for the major structural proteins such as CTCF, RAD21 and the testis-specific CTCF paralog CTCFL at the same cell stages through ChIP-seq/CUT&Tag/CUT&RUN will further strengthen the observations derived from our Micro-C datasets. Lastly, the final developmental stage of spermatogenesis included in our dataset is haploid spermatids. However, after being released from the lumen of the seminiferous tubules into the epididymis, these cells undergo the last stages of maturation, including the gain of motility, to become fully functional sperm. Hence, generating the same dataset for human mature sperm from semen sample will provide a comprehensive understanding of the changes across the entire spermatogenesis process.

# Bibliography


10xGenomics (n.d.). *How does the Gene Expression data compare between whole cells and nuclei?* https://kb.10xgenomics.com/hc/en-us/articles/360054702612-How-does-the-Gene-Expression-data-compare-between-whole-cells-and-nuclei- (cit. on p. 71).

Agarwal, A. et al. (2021). "Male infertility". *The Lancet* 397.10271, pp. 319–333 (cit. on p. 4).

Akgol Oksuz, B. et al. (2021). "Systematic evaluation of chromosome conformation capture assays". *Nature methods* 18.9, pp. 1046–1055 (cit. on pp. 35, 125, 127).

Alavattam, K. G. et al. (2019). "Attenuated chromatin compartmentalization in meiosis and its maturation in sperm development". *Nature structural & molecular biology* 26.3, pp. 175–184 (cit. on pp. 36–38, 115, 126, 129, 135, 137).

Alberts, B. et al. (2002). "Sperm". *Molecular Biology of the Cell. 4th edition*. Garland Science (cit. on p. 99).

Ali, A. et al. (2015). "Prosaposin activates the androgen receptor and potentiates resistance to endocrine treatment in breast cancer". *Breast Cancer Research* 17, pp. 1–19 (cit. on p. 92).

Almiñana, C. et al. (2014). "The battle of the sexes starts in the oviduct: modulation of oviductal transcriptome by X and Y-bearing spermatozoa". *BMC genomics* 15, pp. 1–11 (cit. on p. 100).

Amann, R. P. (2008). "The cycle of the seminiferous epithelium in humans: a need to revisit?" *Journal of andrology* 29.5, pp. 469–487 (cit. on p. 25).

An, N. et al. (2018). "Gene dosage effect of CUX1 in a murine model disrupts HSC homeostasis and controls the severity and mortality of MDS". *Blood, The Journal of the American Society of Hematology* 131.24, pp. 2682–2697 (cit. on pp. 79, 95).

Aras-Tosun, D. (2023). "Genetic alterations in azoospermia patients may reveal potential biomarkers for male infertility: A bioinformatic study." *Journal of Surgery & Medicine (JOSAM)* 7.3 (cit. on p. 77).

Aravin, A. A. et al. (2008). "A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice". *Molecular cell* 31.6, pp. 785–799 (cit. on p. 137).

Azhar, M. et al. (2021). "Towards post-meiotic sperm production: genetic insight into human infertility from mouse models". *International Journal of Biological Sciences* 17.10, p. 2487 (cit. on p. 21).

Baarends, W. M. et al. (1999). "Histone ubiquitination and chromatin remodeling in mouse spermatogenesis". *Developmental biology* 207.2, pp. 322–333 (cit. on p. 21).

Baldi, E. et al. (n.d.). "Jackson Kirkman-Brown, MBE, Ph. D., b, c and other Editorial Board members of the World Health Organization laboratory manual for the examination and processing of human semen" () (cit. on p. 31).

Bao, J. and M. T. Bedford (2016). "Epigenetic regulation of the histone-to-protamine transition during spermiogenesis". *Reproduction (Cambridge, England)* 151.5, R55 (cit. on p. 21).

Barr, M. L. and E. G. Bertram (1949). "A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis". *Nature* 163.4148, pp. 676–677 (cit. on p. 37).

Belaghzal, H. et al. (2017). “Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation”. *Methods* 123, pp. 56–65 (cit. on p. 125).
Belton, J.-M. et al. (2012). “Hi–C: a comprehensive technique to capture the conformation of genomes”. *Methods* 58.3, pp. 268–276 (cit. on p. 35).
Benjamini, Y. and Y. Hochberg (1995). “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. *Journal of the Royal statistical society: series B (Methodological)* 57.1, pp. 289–300 (cit. on p. 49).
Bentsen, M. et al. (2020). “ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation”. *Nature communications* 11.1, p. 4267 (cit. on pp. 32, 47).
Bergen, V., M. Lange, et al. (2020). “Generalizing RNA velocity to transient cell states through dynamical modeling”. *Nature biotechnology* 38.12, pp. 1408–1414 (cit. on pp. 48, 85, 87).
Bergen, V., R. A. Soldatov, et al. (2021). “RNA velocity—current challenges and future perspectives”. *Molecular systems biology* 17.8, e10282 (cit. on p. 86).
Bischof, J. M. et al. (2013). “A genome-wide analysis of open chromatin in human epididymis epithelial cells reveals candidate regulatory elements for genes coordinating epididymal function”. *Biology of reproduction* 89.4, pp. 104–1 (cit. on pp. 77, 94).
Boitrelle, F. et al. (2021). “The sixth edition of the WHO manual for human semen analysis: a critical review and SWOT analysis”. *Life* 11.12, p. 1368 (cit. on p. 16).
Bollschweiler, D. et al. (2019). “Molecular architecture of the SYCP3 fibre and its interaction with DNA”. *Open biology* 9.10, p. 190094 (cit. on p. 18).
Bonfield, J. K. et al. (2021). “HTSlib: C library for reading/writing high-throughput sequencing data”. *Gigascience* 10.2, giab007 (cit. on p. 47).
Boussouar, F. et al. (2008). “A new insight into male genome reprogramming by histone variants and histone code”. *Cell Cycle* 7.22, pp. 3499–3502 (cit. on pp. 21, 115).
Braun, R. E. et al. (1989). “Genetically haploid spermatids are phenotypically diploid”. *Nature* 337.6205, pp. 373–376 (cit. on pp. 64, 99).
Brawand, D. et al. (2011). “The evolution of gene expression levels in mammalian organs”. *Nature* 478.7369, pp. 343–348 (cit. on p. 30).
Brewer, L. R. et al. (1999). “Protamine-induced condensation and decondensation of the same DNA molecule”. *Science* 286.5437, pp. 120–123 (cit. on p. 21).
Bridges, M. C. et al. (2021). “LNCcation: lncRNA localization and function”. *Journal of Cell Biology* 220.2 (cit. on pp. 66, 71).
Britannica, T. (2020). “Editors of encyclopaedia”. *Argon. Encyclopedia Britannica* (cit. on p. 15).
Browaeys, R. et al. (2020). “NicheNet: modeling intercellular communication by linking ligands to target genes”. *Nature methods* 17.2, pp. 159–162 (cit. on p. 98).
Bryant, J. M. et al. (2013). “Separation of spermatogenic cell types using STA-PUT velocity sedimentation”. *JoVE (Journal of Visualized Experiments)* 80, e50648 (cit. on p. 128).
Brykczynska, U. et al. (2010). “Repressive and active histone methylation mark distinct promoters in human and mouse spermatozoa”. *Nature structural & molecular biology* 17.6, pp. 679–687 (cit. on p. 21).
Buenrostro, J. D. et al. (2015). “ATAC-seq: a method for assaying chromatin accessibility genome-wide”. *Current protocols in molecular biology* 109.1, pp. 21–29 (cit. on p. 32).
Caldeira-Brant, A. L. et al. (2020). “A subpopulation of human Adark spermatogonia behaves as the reserve stem cell”. *Reproduction* 159.4, pp. 437–451 (cit. on pp. 16, 72).
Cardoso-Moreira, M. et al. (2019). “Gene expression across mammalian organ development”. *Nature* 571.7766, pp. 505–509 (cit. on p. 30).
Carmell, M. A. et al. (2007). “MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline”. *Developmental cell* 12.4, pp. 503–514 (cit. on p. 137).
Carone, B. R. et al. (2014). “High-resolution mapping of chromatin packaging in mouse embryonic stem cells and sperm”. *Developmental cell* 30.1, pp. 11–22 (cit. on p. 21).

Carrieri, C. et al. (2017). “A transit-amplifying population underpins the efficient regenerative capacity of the testis”. *Journal of Experimental Medicine* 214.6, pp. 1631–1641 (cit. on pp. 30, 136).

Carvalho, J. O. et al. (2013). “Nanoscale differences in the shape and size of X and Y chromosome-bearing bovine sperm heads assessed by atomic force microscopy”. *PloS one* 8.3, e59387 (cit. on p. 100).

Chain, M. (2017). *Definition of Markov Chain in US English by Oxford Dictionaries* (cit. on p. 86).

Chapman, K. M. et al. (2015). “NRG1 and KITL signal downstream of retinoic acid in the germline to support soma-free syncytial growth of differentiating spermatogonia”. *Cell death discovery* 1.1, pp. 1–13 (cit. on p. 77).

Chen, H., E. Murray, et al. (2021). “Dissecting mammalian spermatogenesis using spatial transcriptomics”. *Cell reports* 37.5 (cit. on pp. 28, 71).

Chen, H., M.-P. Scott-Boyer, et al. (2022). “Sperm Heterogeneity Accounts for Sperm DNA Methylation Variations Observed in the Caput Epididymis, Independently From DNMT/TET Activities”. *Frontiers in Cell and Developmental Biology*, p. 618 (cit. on p. 38).

Chen, X., Y. Yue, et al. (2014). “Identification and characterization of genes differentially expressed in X and Y sperm using suppression subtractive hybridization and cDNA microarray”. *Molecular Reproduction and Development* 81.10, pp. 908–917 (cit. on p. 100).

Chen, X., H. Zhu, et al. (2012). “Identification of differentially expressed proteins between bull X and Y spermatozoa”. *Journal of proteomics* 77, pp. 59–67 (cit. on p. 100).

Chen, X., Y. Ke, et al. (2019). “Key role for CTCF in establishing chromatin structure in human embryos”. *Nature* 576.7786, pp. 306–310 (cit. on pp. 38, 115, 130, 137).

Cheng, X. and Q. Sun (2019). “RUBCNL/Pacer and RUBCN/Rubicon in regulation of autolysosome formation and lipid metabolism”. *Autophagy* 15.6, pp. 1120–1121 (cit. on p. 31).

Choy, B. et al. (2016). “MCM4 and MCM7, potential novel proliferation markers, significantly correlated with Ki-67, Bmi1, and cyclin E expression in esophageal adenocarcinoma, squamous cell carcinoma, and precancerous lesions”. *Human pathology* 57, pp. 126–135 (cit. on p. 16).

Christensen, A. K. (2007). “A history of Leydig cell research”. *The Leydig cell in health and disease*, pp. 3–30 (cit. on p. 23).

Ciccodicola, A. et al. (2000). “Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region”. *Human Molecular Genetics* 9.3, pp. 395–401 (cit. on pp. 19, 101).

Clermont, Y. (1963). “The cycle of the seminiferous epithelium in man”. *American Journal of Anatomy* 112.1, pp. 35–51 (cit. on pp. 24, 25).

– (1966). “Spermatogenesis in man: a study of the spermatogonial population”. *Fertility and sterility* 17.6, pp. 705–721 (cit. on p. 24).

– (1969). “Two classes of spermatogonial stem cells in the monkey (Cercopithecus aethiops)”. *American Journal of Anatomy* 126.1, pp. 57–71 (cit. on pp. 16, 72).

– (1972). “Kinetics of spermatogenesis in mammals: seminiferous epithelium cycle and spermatogonial renewal.” *Physiological reviews* 52.1, pp. 198–236 (cit. on pp. 16, 72).

Cremer, T. and M. Cremer (2010). “Chromosome territories”. *Cold Spring Harbor perspectives in biology* 2.3, a003889 (cit. on p. 33).

Crichton, J. H. et al. (2023). “Structural maturation of SYCP1-mediated meiotic chromosome synapsis by SYCE3”. *Nature structural & molecular biology* 30.2, pp. 188–199 (cit. on p. 18).

Daish, T. J. et al. (2015). “Lack of sex chromosome specific meiotic silencing in platypus reveals origin of MSCI in therian mammals”. *BMC biology* 13.1, pp. 1–13 (cit. on p. 31).

Danecek, P. et al. (2021). *Twelve years of SAMtools and BCFtools. Gigascience 10: giab008* (cit. on p. 47).

Darrow, E. M. et al. (2016). “Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture”. *Proceedings of the National Academy of Sciences* 113.31, E4504–E4512 (cit. on p. 37).

Davis, M. P. et al. (2017). “Transposon-driven transcription is a conserved feature of vertebrate spermatogenesis and transcript evolution”. *EMBO reports* 18.7, pp. 1231–1247 (cit. on p. 30).

De Canio, M. et al. (2014). “Differential protein profile in sexed bovine semen: shotgun proteomics investigation”. *Molecular BioSystems* 10.6, pp. 1264–1271 (cit. on p. 100).

Delihas, N. (2018). “A family of long intergenic non-coding RNA genes in human chromosomal region 22q11. 2 carry a DNA translocation breakpoint/AT-rich sequence”. *PLoS One* 13.4, e0195702 (cit. on p. 113).

– (2020). “Formation of human long intergenic non-coding RNA genes, pseudogenes, and protein genes: Ancestral sequences are key players”. *PLoS One* 15.3, e0230236 (cit. on p. 113).

Dettin, L. et al. (2003). “Morphological characterization of the spermatogonial subtypes in the neonatal mouse testis”. *Biology of Reproduction* 69.5, pp. 1565–1571 (cit. on p. 135).

Di Giacomo, M. et al. (2013). “Multiple epigenetic mechanisms and the piRNA pathway enforce LINE1 silencing during adult spermatogenesis”. *Molecular cell* 50.4, pp. 601–608 (cit. on pp. 18, 19).

Di Persio, S. and N. Neuhaus (2023). “Human spermatogonial stem cells and their niche in male (in) fertility: novel concepts from single-cell RNA-sequencing”. *Human Reproduction* 38.1, pp. 1–13 (cit. on pp. 28, 43, 45, 73).

Di Persio, S., R. Saracino, et al. (2017). “Spermatogonial kinetics in humans”. *Development* 144.19, pp. 3430–3439 (cit. on pp. 94, 95, 133).

Di Persio, S., T. Tekath, et al. (2021). “Single-cell RNA-seq unravels alterations of the human spermatogonial stem cell compartment in patients with impaired spermatogenesis”. *Cell Reports Medicine* 2.9, p. 100395 (cit. on pp. 22, 28, 31, 43–45, 54, 62, 64–67, 72, 73, 94, 96, 100, 133, 135).

Díaz, N. et al. (2018). “Chromatin conformation analysis of primary patient tissue using a low input Hi-C method”. *Nature communications* 9.1, p. 4938 (cit. on p. 56).

Dimitrov, D. et al. (2022). “Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data”. *Nature communications* 13.1, p. 3224 (cit. on pp. 49, 88).

Dixon, J. R. et al. (2012). “Topological domains in mammalian genomes identified by analysis of chromatin interactions”. *Nature* 485.7398, pp. 376–380 (cit. on p. 33).

Dominko, T. and N. First (1997). “Relationship between the maturational state of oocytes at the time of insemination and sex ratio of subsequent early bovine embryos”. *Theriogenology* 47.5, pp. 1041–1050 (cit. on p. 100).

Durand, N. C. et al. (2016). “Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments”. *Cell systems* 3.1, pp. 95–98 (cit. on p. 35).

Dym, M. and D. W. Fawcett (1971). “Further observations on the numbers of spermatogonia, spermatocytes, and spermatids connected by intercellular bridges in the mammalian testis”. *Biology of reproduction* 4.2, pp. 195–215 (cit. on pp. 16, 99).

Earnshaw, W. C. and N. Rothfield (1985). “Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma”. *Chromosoma* 91.3-4, pp. 313–321 (cit. on p. 18).

Eddy, E. (2006). “Knobil and Neill’s physiology of reproduction”. *(No Title)*, p. 3 (cit. on p. 21).

Efremova, M. et al. (2020). “CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes”. *Nature protocols* 15.4, pp. 1484–1506 (cit. on pp. 48, 98).

Engreitz, J. M. et al. (2013). “The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome”. *Science* 341.6147, p. 1237973 (cit. on p. 33).

Ernst, C. et al. (2023). “Efficient and sensitive profiling of RNA–protein interactions using TLC-CLIP”. *Nucleic Acids Research* 51.13, e70–e70 (cit. on p. 136).

Ewen, K. A. et al. (2013). “Expression of FGFR3 during human testis development and in germ cell-derived tumours of young adults”. *International Journal of Developmental Biology* 57.2-3-4, pp. 141–151 (cit. on p. 82).

Fagerberg, L. et al. (2014). “Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics”. *Molecular & cellular proteomics* 13.2, pp. 397–406 (cit. on p. 113).

Fayomi, A. P. and K. E. Orwig (2018). “Spermatogonial stem cells and spermatogenesis in mice, monkeys and men”. *Stem cell research* 29, pp. 207–214 (cit. on p. 135).

Finak, G. et al. (2015). “MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data”. *Genome biology* 16.1, pp. 1–13 (cit. on p. 45).

Flaquer, A. et al. (2008). “The human pseudoautosomal regions: a review for genetic epidemiologists”. *European Journal of Human Genetics* 16.7, pp. 771–779 (cit. on p. 101).

França, L. R. et al. (2016). “The Sertoli cell: one hundred fifty years of beauty and plasticity”. *Andrology* 4.2, pp. 189–212 (cit. on p. 22).

Fujishiro, S. and M. Sasai (2022). “Generation of dynamic three-dimensional genome structure through phase separation of chromatin”. *Proceedings of the National Academy of Sciences* 119.22, e2109838119 (cit. on p. 33).

Fukuda, K., Y. Makino, et al. (2022). “Potential role of KRAB-ZFP binding and transcriptional states on DNA methylation of retroelements in human male germ cells”. *Elife* 11, e76822 (cit. on pp. 4, 101, 113, 134, 137).

Fukuda, K. and Y. Shinkai (2020). “SETDB1-mediated silencing of retroelements”. *Viruses* 12.6, p. 596 (cit. on p. 113).

Fullard, J. F. et al. (2018). “An atlas of chromatin accessibility in the adult human brain”. *Genome research* 28.8, pp. 1243–1252 (cit. on p. 32).

Galan, C. et al. (2021). “Stability of the cytosine methylome during post-testicular sperm maturation in mouse”. *PLoS genetics* 17.3, e1009416 (cit. on p. 38).

Garcia-Alonso, L. et al. (2022). “Single-cell roadmap of human gonadal development”. *Nature* 607.7919, pp. 540–547 (cit. on pp. 29, 54).

Gatewood, J. et al. (1987). “Sequence-specific packaging of DNA in human sperm chromatin”. *Science* 236.4804, pp. 962–964 (cit. on p. 21).

Geng, K. (2020). “Post-translational modifications of the ligands: Requirement for TAM receptor activation”. *International Review of Cell and Molecular Biology* 357, pp. 35–55 (cit. on p. 98).

Geremia, R. et al. (1977). “RNA synthesis in spermatocytes and spermatids and preservation of meiotic RNA during spermiogenesis in the mouse”. *Cell differentiation* 5.5-6, pp. 343–355 (cit. on pp. 71, 88, 114).

Germain, P.-L. et al. (2021). “Doublet identification in single-cell sequencing data using scDblFinder”. *F1000Research* 10 (cit. on p. 42).

Ghoreishifar, M. et al. (2023). “Genome-wide assessment and mapping of inbreeding depression identifies candidate genes associated with semen traits in Holstein bulls”. *BMC genomics* 24.1, pp. 1–11 (cit. on p. 77).

Gibcus, J. H. et al. (2018). “A pathway for mitotic chromosome formation”. *Science* 359.6376, eaao6135 (cit. on pp. 127, 130).

Gilbert, S. F. and M. Barresi (2000). “Developmental biology, ed”. *Sunderland (MA): Sinauer* (cit. on pp. 16, 99).

Gkountela, S. et al. (2013). “The ontogeny of cKIT+ human primordial germ cells proves to be a resource for human germ line reprogramming, imprint erasure and in vitro differentiation”. *Nature cell biology* 15.1, pp. 113–122 (cit. on p. 137).

Goodier, J. L. (2016). “Restricting retrotransposons: a review”. *Mobile DNA* 7, pp. 1–30 (cit. on p. 113).

Grabski, I. N. et al. (2023). “Significance analysis for clustering with single-Cell RNA-sequencing data”. *Nature Methods*, pp. 1–7 (cit. on p. 94).

Granja, J. M. et al. (2021). “ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis”. *Nature genetics* 53.3, pp. 403–411 (cit. on p. 46).

Grimes Jr, S. R. and N. Henderson (1984). “Hyperacetylation of histone H4 in rat testis spermatids”. *Experimental cell research* 152.1, pp. 91–97 (cit. on p. 21).

Guo, F., L. Yan, et al. (2015). “The transcriptome and DNA methylome landscapes of human primordial germ cells”. *Cell* 161.6, pp. 1437–1452 (cit. on p. 137).

Guo, J., E. J. Grow, H. Mlcochova, et al. (2018). “The adult human testis transcriptional cell atlas”. *Cell research* 28.12, pp. 1141–1157 (cit. on pp. 27, 28, 31, 32, 43, 62, 64, 65, 67, 70, 72, 73, 94–96, 100, 114, 133, 135–137).

Guo, J., E. J. Grow, C. Yi, et al. (2017). “Chromatin and single-cell RNA-seq profiling reveal dynamic signaling and metabolic transitions during human spermatogonial stem cell development”. *Cell stem cell* 21.4, pp. 533–546 (cit. on pp. 64, 72, 100).

Guo, J., X. Nie, et al. (2020). “The dynamic transcriptional cell atlas of testis development during human puberty”. *Cell stem cell* 26.2, pp. 262–276 (cit. on pp. 28–31, 54, 72, 100, 137).

Guo, J., E. Sosa, et al. (2021). “Single-cell analysis of the developing human testis reveals somatic niche cell specification and fetal germline stem cell establishment”. *Cell Stem Cell* 28.4, pp. 764–778 (cit. on pp. 28–30, 54, 64, 72, 100).

Hafner, M. et al. (2021). “CLIP and complementary methods”. *Nature Reviews Methods Primers* 1.1, p. 20 (cit. on p. 136).

Hammoud, S. S. et al. (2009). “Distinctive chromatin in human sperm packages genes for embryo development”. *Nature* 460.7254, pp. 473–478 (cit. on p. 22).

Hanawalt, P. C. and G. Spivak (2008). “Transcription-coupled DNA repair: two decades of progress and surprises”. *Nature reviews Molecular cell biology* 9.12, pp. 958–970 (cit. on p. 30).

Handel, M. A. (2004). “The XY body: a specialized meiotic chromatin domain”. *Experimental cell research* 296.1, pp. 57–63 (cit. on p. 19).

Hao, Y. et al. (2021). “Integrated analysis of multimodal single-cell data”. *Cell* 184.13, pp. 3573–3587 (cit. on p. 41).

He, S. et al. (2021). “High-plex multiomic analysis in FFPE tissue at single-cellular and subcellular resolution by spatial molecular imaging”. *bioRxiv*, pp. 2021–11 (cit. on p. 27).

Heijden, G. W. van der et al. (2005). “Asymmetry in histone H3 variants and lysine methylation between paternal and maternal chromatin of the early mouse zygote”. *Mechanisms of development* 122.9, pp. 1008–1022 (cit. on p. 21).

Heinz, S. et al. (2018). “Transcription elongation can affect genome 3D structure”. *Cell* 174.6, pp. 1522–1536 (cit. on p. 124).

Helena Mangs, A. and B. J. Morris (2007). “The human pseudoautosomal region (PAR): origin, function and future”. *Current genomics* 8.2, pp. 129–136 (cit. on p. 101).

Heller, G. (1964). “Kinetics of the germinal epithelium in man”. *Recent Prog Horm Res* 20, pp. 545–575 (cit. on pp. 16, 96).

Hermann, B. P. et al. (2018). “The mammalian spermatogenesis single-cell transcriptome, from spermatogonial stem cells to spermatids”. *Cell reports* 25.6, pp. 1650–1667 (cit. on pp. 27, 64, 70, 100).

Hinch, A. G. et al. (2020). “The configuration of RPA, RAD51, and DMC1 binding in meiosis reveals the nature of critical recombination intermediates”. *Molecular cell* 79.4, pp. 689–701 (cit. on p. 18).

Hisano, M. et al. (2013). “Genome-wide chromatin analysis in mature mouse and human spermatozoa”. *Nature protocols* 8.12, pp. 2449–2470 (cit. on p. 38).

Hogarth, C. A. et al. (2010). “EGR4 displays both a cell-and intracellular-specific localization pattern in the developing murine testis”. *Developmental Dynamics* 239.11, pp. 3106–3114 (cit. on p. 31).

Holt, W. V. and A. Fazeli (2010). “The oviduct as a complex mediator of mammalian sperm function and selection”. *Molecular reproduction and development* 77.11, pp. 934–943 (cit. on p. 100).

Howards, S. S. (1997). “Antoine van Leeuwenhoek and the discovery of sperm”. *Fertility and sterility* 1.67, pp. 16–17 (cit. on p. 23).

Hsieh, T.-H. S., C. Cattoglio, et al. (2020). “Resolving the 3D landscape of transcription-linked mammalian chromatin folding”. *Molecular cell* 78.3, pp. 539–553 (cit. on pp. 35, 51, 127, 130).

Hsieh, T.-H. S., G. Fudenberg, et al. (2016). “Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome”. *Nature methods* 13.12, pp. 1009–1011 (cit. on p. 123).

*https://wp.unil.ch/gtf/single-cell-experiments-10x/* (n.d.) (cit. on p. 26).

*https://www.10xgenomics.com/* (n.d.) (cit. on pp. 26, 32).

*https://www.illumina.com/* (n.d.) (cit. on p. 26).

*https://www.thermofisher.com/order/fluorescence-spectraviewer/* (n.d.) (cit. on p. 117).

Huang, C. R. L. et al. (2012). “Active transposition in genomes”. *Annual review of genetics* 46, pp. 651–675 (cit. on p. 113).

Huckins, C. (1971). “The spermatogonial stem cell population in adult rats. I. Their morphology, proliferation and maturation”. *The Anatomical Record* 169.3, pp. 533–557 (cit. on pp. 64, 99).

Hug, C. B. and J. M. Vaquerizas (2018). “The birth of the 3D genome during early embryonic development”. *Trends in Genetics* 34.12, pp. 903–914 (cit. on p. 33).

Imbeault, M. et al. (2017). “KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks”. *Nature* 543.7646, pp. 550–554 (cit. on p. 31).

Iyer, K. S. et al. (2023). “Cellular compartmentalisation and receptor promiscuity as a strategy for accurate and robust inference of position during morphogenesis”. *Elife* 12, e79257 (cit. on p. 98).

Jan, S. Z. et al. (2017). “Unraveling transcriptome dynamics in human spermatogenesis”. *Development* 144.20, pp. 3659–3673 (cit. on pp. 28, 72).

Jin, F. et al. (2013). “A high-resolution map of the three-dimensional chromatin interactome in human cells”. *Nature* 503.7475, pp. 290–294 (cit. on p. 33).

Jin, S. et al. (2021). “Inference and analysis of cell-cell communication using CellChat”. *Nature communications* 12.1, p. 1088 (cit. on pp. 97, 98).

Jodar, M. et al. (2013). “The presence, role and clinical use of spermatozoal RNAs”. *Human reproduction update* 19.6, pp. 604–624 (cit. on p. 100).

Johnson, G. D. et al. (2011). “The sperm nucleus: chromatin, RNA and the nuclear matrix”. *Reproduction (Cambridge, England)* 141.1, p. 21 (cit. on p. 88).

Johnsson, K. et al. (2014). “Low bias local intrinsic dimension estimation from expected simplex skewness”. *IEEE transactions on pattern analysis and machine intelligence* 37.1, pp. 196–202 (cit. on p. 94).

Kalhor, R. et al. (2012). “Genome architectures revealed by tethered chromosome conformation capture and population-based modeling”. *Nature biotechnology* 30.1, pp. 90–98 (cit. on p. 33).

Kaya-Okur, H. S. et al. (2019). “CUT&Tag for efficient epigenomic profiling of small samples and single cells”. *Nature communications* 10.1, p. 1930 (cit. on p. 56).

Keeney, S. (2007). “Spo11 and the formation of DNA double-strand breaks in meiosis”. *Recombination and meiosis: crossing-over and disjunction*. Springer, pp. 81–123 (cit. on pp. 18, 19).

Kempfer, R. and A. Pombo (2020). “Methods for mapping 3D chromosome architecture”. *Nature Reviews Genetics* 21.4, pp. 207–226 (cit. on p. 34).

Kerr, J. and D. De Kretser (1974). “Proceedings: The role of the Sertoli cell in phagocytosis of the residual bodies of spermatids”. *Journal of reproduction and fertility* 36.2, pp. 439–440 (cit. on p. 21).

Kitadate, Y. et al. (2019). “Competition for mitogens regulates spermatogenic stem cell homeostasis in an open niche”. *Cell Stem Cell* 24.1, pp. 79–92 (cit. on p. 93).

Klein, A. M. et al. (2015). “Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells”. *Cell* 161.5, pp. 1187–1201 (cit. on p. 26).

Knight, P. A. and D. Ruiz (2013). “A fast algorithm for matrix balancing”. *IMA Journal of Numerical Analysis* 33.3, pp. 1029–1047 (cit. on p. 53).

Kossack, N. et al. (2013). “A combined approach facilitates the reliable detection of human spermatogonia in vitro”. *Human Reproduction* 28.11, pp. 3012–3025 (cit. on pp. 40, 55).

Kotaja, N. and P. Sassone-Corsi (2007). “The chromatoid body: a germ-cell-specific RNA-processing centre”. *Nature reviews Molecular cell biology* 8.1, pp. 85–90 (cit. on pp. 66, 71).

Krawetz, S. A. (2005). “Paternal contribution: new insights and future challenges”. *Nature Reviews Genetics* 6.8, pp. 633–642 (cit. on p. 100).

Krietenstein, N. et al. (2020). “Ultrastructural details of mammalian chromosome architecture”. *Molecular cell* 78.3, pp. 554–565 (cit. on pp. 35, 123).

Kruse, K. et al. (2020). “FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data”. *Genome biology* 21.1, pp. 1–19 (cit. on p. 53).

La Manno, G. et al. (2018). “RNA velocity of single cells”. *Nature* 560.7719, pp. 494–498 (cit. on pp. 48, 85, 87, 96).

Lajoie, B. R. et al. (2015). “The Hitchhiker’s guide to Hi-C analysis: practical guidelines”. *Methods* 72, pp. 65–75 (cit. on pp. 124, 125).

Landry, C. et al. (2000). “Monoclonal antibody 57B stains tumor tissues that express gene MAGE-A4”. *International journal of cancer* 86.6, pp. 835–841 (cit. on pp. 73, 118).

Lareau, C. A. et al. (2021). “Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling”. *Nature Biotechnology* 39.4, pp. 451–461 (cit. on p. 70).

Larney, C. et al. (2014). “Switching on sex: transcriptional regulation of the testis-determining gene Sry”. *Development* 141.11, pp. 2195–2205 (cit. on p. 99).

Lau, X. et al. (2020). “Single-cell RNA sequencing of the cynomolgus macaque testis reveals conserved transcriptional profiles during mammalian spermatogenesis”. *Developmental Cell* 54.4, pp. 548–566 (cit. on p. 30).

Laurentino, S. et al. (2019). “High-resolution analysis of germ cells from men with sex chromosomal aneuploidies reveals normal transcriptome but impaired imprinting”. *Clinical epigenetics* 11.1, pp. 1–13 (cit. on pp. 40, 55).

Leeuwenhoek, A. v. and G. v. Rijnberk (1939). “Alle de brieven van Antoni van Leeuwenhoek: uitgegeven, geıHERE!HERE!llustreerd en van aanteekeningen voorzien door een commissie van nederlandsche geleerden...” *(No Title)* (cit. on p. 23).

Lemke, G. (2013). “Biology of the TAM receptors”. *Cold Spring Harbor perspectives in biology* 5.11, a009076 (cit. on pp. 93, 97).

Leydig, F. (1850). “On the anatomy of the male sex organs and anal glands of mammals”. *Z Science Zool* 2, p. 1 (cit. on p. 23).

Li, C., M. C. Virgilio, et al. (2023). “Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction”. *Nature Biotechnology* 41.3, pp. 387–398 (cit. on pp. 48, 85, 96).

Li, H. and R. Durbin (2009). “Fast and accurate short read alignment with Burrows–Wheeler transform”. *bioinformatics* 25.14, pp. 1754–1760 (cit. on p. 53).

Lieberman-Aiden, E. et al. (2009). “Comprehensive mapping of long-range interactions reveals folding principles of the human genome”. *science* 326.5950, pp. 289–293 (cit. on pp. 33–35).

Ling, X. et al. (2022). “The dynamics of three-dimensional chromatin organization and phase separation in cell fate transitions and diseases”. *Cell Regeneration* 11.1, p. 42 (cit. on p. 33).

Liu, S. et al. (2021). “Role of RNA N6-methyladenosine modification in male infertility and genital system tumors”. *Frontiers in cell and developmental biology* 9, p. 676364 (cit. on pp. 71, 136).

Liu, Y. et al. (2015). “Fractionation of human spermatogenic cells using STA-PUT gravity sedimentation and their miRNA profiling”. *Scientific reports* 5.1, pp. 1–11 (cit. on p. 128).

Loukinov, D. I. et al. (2002). “BORIS, a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma”. *Proceedings of the National Academy of Sciences* 99.10, pp. 6806–6811 (cit. on p. 130).

Lu, Q., M. Gore, et al. (1999). “Tyro-3 family receptors are essential regulators of mammalian spermatogenesis”. *Nature* 398.6729, pp. 723–728 (cit. on pp. 93, 97).

Lu, S., C. Zong, et al. (2012). “Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing”. *science* 338.6114, pp. 1627–1630 (cit. on p. 20).

Lun, A. T. et al. (2016). “A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor”. *F1000Research* 5 (cit. on p. 42).

Luo, Z. et al. (2020). “Reorganized 3D genome structures support transcriptional regulation in mouse spermatogenesis”. *Iscience* 23.4, p. 101034 (cit. on pp. 36–38, 115, 126, 129, 130, 135).

Lupiáñez, D. G. et al. (2015). “Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions”. *Cell* 161.5, pp. 1012–1025 (cit. on p. 33).

Macosko, E. Z. et al. (2015). “Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets”. *Cell* 161.5, pp. 1202–1214 (cit. on p. 26).

Maekawa, M. et al. (1996). “Peritubular myoid cells in the testis: their structure and function”. *Archives of histology and cytology* 59.1, pp. 1–13 (cit. on p. 31).

Maksakova, I. A. et al. (2013). “Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells”. *Epigenetics & chromatin* 6, pp. 1–16 (cit. on p. 113).

Marcon, L. and G. Boissonneault (2004). “Transient DNA strand breaks during mouse and human spermiogenesis: new insights in stage specificity and link to chromatin remodeling”. *Biology of reproduction* 70.4, pp. 910–918 (cit. on p. 21).

Mayerhofer, A. (2013). “Human testicular peritubular cells: more than meets the eye”. *Reproduction* 145.5, R107–R116 (cit. on pp. 22, 23).

McCarthy, D. J. et al. (2017). “Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R”. *Bioinformatics* 33.8, pp. 1179–1186 (cit. on p. 42).

McLachlan, R. I. et al. (2007). “Histological evaluation of the human testis—approaches to optimizing the clinical value of the assessment: mini review”. *Human reproduction* 22.1, pp. 2–16 (cit. on pp. 24, 25).

Meistrich, M. L., B. Mohapatra, et al. (2003). “Roles of transition nuclear proteins in spermiogenesis”. *Chromosoma* 111, pp. 483–488 (cit. on p. 21).

Meistrich, M. L., P. K. Trostle-Weige, et al. (1992). “Highly acetylated H4 is associated with histone displacement in rat spermatids”. *Molecular reproduction and development* 31.3, pp. 170–181 (cit. on p. 21).

Merritt, C. R. et al. (2020). “Multiplex digital spatial profiling of proteins and RNA in fixed tissue”. *Nature biotechnology* 38.5, pp. 586–599 (cit. on p. 27).

Meyer, C. et al. (2017). “Simultaneous detection of the subcellular localization of RNAs and proteins in cultured cells by combined multicolor RNA-FISH and IF”. *Methods* 118, pp. 101–110 (cit. on p. 113).

Miura, M. and H. Chen (2020). “CUT&RUN detects distinct DNA footprints of RNA polymerase II near the transcription start sites”. *Chromosome Research* 28, pp. 381–393 (cit. on p. 129).

Monesi, V. (1964). “Ribonucleic acid synthesis during mitosis and meiosis in the mouse testis”. *The Journal of cell biology* 22.3, pp. 521–532 (cit. on pp. 17, 19, 130).

Morales, C. R. and H. Badran (2003). “Prosaposin ablation inactivates the MAPK and Akt signaling pathways and interferes with the development of the prostate gland”. *Asian journal of andrology* 5.1, pp. 57–64 (cit. on p. 92).

Morales, C. R., X. Q. Wu, et al. (1998). “The DNA/RNA-binding protein, TB-RBP, moves from the nucleus to the cytoplasm and through intercellular bridges in male germ cells”. *Developmental biology* 201.1, pp. 113–123 (cit. on pp. 64, 99).

Morales, C. R., Q. Zhao, et al. (2000). “Targeted disruption of the mouse prosaposin gene affects the development of the prostate gland and other male reproductive organs”. *Journal of andrology* 21.6, pp. 765–775 (cit. on p. 92).

Morgan, M., Y. Kabayama, et al. (2019). “A programmed wave of uridylation-primed mRNA degradation is essential for meiotic progression and mammalian spermatogenesis”. *Cell research* 29.3, pp. 221–232 (cit. on p. 71).

Morgan, M., L. Kumar, et al. (2021). “Post-transcriptional regulation in spermatogenesis: all RNA pathways lead to healthy sperm”. *Cellular and Molecular Life Sciences* 78, pp. 8049–8071 (cit. on p. 71).

Mruk, D. D. and C. Y. Cheng (2004). “Sertoli-Sertoli and Sertoli-germ cell interactions and their significance in germ cell movement in the seminiferous epithelium during spermatogenesis”. *Endocrine reviews* 25.5, pp. 747–806 (cit. on p. 22).

Mueller, J. L. et al. (2008). “The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression”. *Nature genetics* 40.6, pp. 794–799 (cit. on p. 136).

Murat, F. et al. (2023). “The molecular evolution of spermatogenesis across mammals”. *Nature* 613.7943, pp. 308–316 (cit. on pp. 28, 30, 31, 54, 64, 70, 100, 102–104, 110, 113, 136).

Nakagawa, T. et al. (2021). “A multistate stem cell dynamics maintains homeostasis in mouse spermatogenesis”. *Cell Reports* 37.3 (cit. on pp. 30, 84, 97, 136).

Namekawa, S. H. et al. (2006). “Postmeiotic sex chromatin in the male germline of mice”. *Current Biology* 16.7, pp. 660–667 (cit. on pp. 19, 136).

Naumova, N. et al. (2013). “Organization of the mitotic chromosome”. *Science* 342.6161, pp. 948–953 (cit. on pp. 127, 130).

Neuhaus, N. et al. (2017). “Single-cell gene expression analysis reveals diversity among human spermatogonia”. *MHR: Basic science of reproductive medicine* 23.2, pp. 79–90 (cit. on pp. 40, 55).

Nie, X. et al. (2022). “Single-cell analysis of human testis aging and correlation with elevated body mass index”. *Developmental Cell* 57.9, pp. 1160–1176 (cit. on pp. 29, 30, 54).

Nishana, M. et al. (2020). “Defining the relative and combined contribution of CTCF and CTCFL to genomic regulation”. *Genome biology* 21, pp. 1–34 (cit. on p. 130).

Nora, E. P. et al. (2012). “Spatial partitioning of the regulatory landscape of the X-inactivation centre”. *Nature* 485.7398, pp. 381–385 (cit. on p. 33).

O’Donnell, L. et al. (2011). “Spermiation: the process of sperm release”. *Spermatogenesis* 1.1, pp. 14–35 (cit. on p. 22).

O'Connor, C. (2008). "Meiosis, genetic recombination, and sexual reproduction". *Nature education* 1.1, p. 174 (cit. on p. 115).

O'Donnell, L. (2014). "Mechanisms of spermiogenesis and spermiation and how they are disturbed". *Spermatogenesis* 4.2, e979623 (cit. on p. 99).

Oatley, M. J. et al. (2011). "Sertoli cells dictate spermatogonial stem cell niches in the mouse testis". *Biology of reproduction* 84.4, pp. 639–645 (cit. on p. 22).

Organization, W. H. et al. (2023). "Infertility prevalence estimates: 1990–2021" (cit. on p. 4).

Ostertag, E. M. et al. (2003). "SVA elements are nonautonomous retrotransposons that cause disease in humans". *The American Journal of Human Genetics* 73.6, pp. 1444–1451 (cit. on p. 113).

Padmanabhan, V. et al. (2004). "DNA replication regulation protein Mcm7 as a marker of proliferation in prostate cancer". *Journal of clinical pathology* 57.10, pp. 1057–1062 (cit. on p. 16).

Paronetto, M. P. et al. (2011). "Sam68 marks the transcriptionally active stages of spermatogenesis and modulates alternative splicing in male germ cells". *Nucleic acids research* 39.12, pp. 4961–4974 (cit. on pp. 17, 19, 130).

Patel, L. et al. (2019). "Dynamic reorganization of the genome shapes the recombination landscape in meiotic prophase". *Nature structural & molecular biology* 26.3, pp. 164–174 (cit. on pp. 36–38, 115, 126, 129, 130, 135).

Patriarca, C. et al. (2019). "Enrico Sertoli and the supporting cells of the testis "Morphology is function"". *Pathologica* 111.4, p. 375 (cit. on p. 23).

Patwardhan, A. et al. (2021). "Post-translational modifications of G Protein–Coupled receptors control cellular signaling dynamics in space and time". *Pharmacological reviews* 73.1, pp. 120–151 (cit. on p. 98).

Pękowska, A. et al. (2018). "Gain of CTCF-anchored chromatin loops marks the exit from naive pluripotency". *Cell Systems* 7.5, pp. 482–495 (cit. on p. 33).

Perrard, M.-H. et al. (2016). "Complete human and rat ex vivo spermatogenesis from fresh or frozen testicular tissue". *Biology of reproduction* 95.4, pp. 89–1 (cit. on p. 97).

Pessot, C. A. et al. (1989). "Presence of RNA in the sperm nucleus". *Biochemical and biophysical research communications* 158.1, pp. 272–278 (cit. on p. 100).

Peters, J.-M. and T. Nishiyama (2012). "Sister chromatid cohesion". *Cold Spring Harbor Perspectives in Biology* 4.11, a011130 (cit. on p. 17).

Qiao, H. et al. (2012). "Interplay between synaptonemal complex, homologous recombination, and centromeres during mammalian meiosis". *PLoS genetics* 8.6, e1002790 (cit. on p. 18).

Qiu, X. et al. (2017). "Reversed graph embedding resolves complex single-cell trajectories". *Nature methods* 14.10, pp. 979–982 (cit. on p. 96).

Rahman, M. S. and M.-G. Pang (2020). "New biological insights on X and Y chromosome-bearing spermatozoa". *Frontiers in cell and developmental biology* 7, p. 388 (cit. on p. 100).

Rajachandran, S. et al. (2023). "Dissecting the spermatogonial stem cell niche using spatial transcriptomics". *Cell Reports* 42.7 (cit. on pp. 28, 72, 97, 98).

Ramm, S. A. et al. (2014). "Sperm competition and the evolution of spermatogenesis". *Molecular human reproduction* 20.12, pp. 1169–1179 (cit. on p. 30).

Rao, S. S., M. H. Huntley, et al. (2014). "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping". *Cell* 159.7, pp. 1665–1680 (cit. on p. 35).

Rao, V. S., K. Srinivas, et al. (2014). "Protein-protein interaction detection: methods and analysis". *International journal of proteomics* 2014 (cit. on p. 98).

Rathke, C. et al. (2014). "Chromatin dynamics during spermiogenesis". *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1839.3, pp. 155–168 (cit. on p. 21).

Raverdeau, M. et al. (2012). "Retinoic acid induces Sertoli cell paracrine signals for spermatogonia differentiation but cell autonomously drives spermatocyte meiosis". *Proceedings of the National Academy of Sciences* 109.41, pp. 16582–16587 (cit. on p. 22).

Reece, J. B. et al. (2015). *Campbell Biology Australian and New Zealand Edition*. Vol. 10. Pearson Higher Education AU (cit. on p. 17).

Rodriques, S. G. et al. (2019). "Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution". *Science* 363.6434, pp. 1463–1467 (cit. on pp. 26, 27).

Rolland, A. D. et al. (2013). "Identification of genital tract markers in the human seminal plasma using an integrative genomics approach". *Human reproduction* 28.1, pp. 199–209 (cit. on p. 77).

Rousseaux, S. et al. (2008). "Epigenetic reprogramming of the male genome during gametogenesis and in the zygote". *Reproductive biomedicine online* 16.4, pp. 492–503 (cit. on pp. 21, 115).

Sarkar, S. et al. (1984). "Swimming behavior of X and Y human sperm". *Differentiation* 27.1-3, pp. 120–125 (cit. on p. 100).

Schep, A. N. et al. (2017). "chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data". *Nature methods* 14.10, pp. 975–978 (cit. on p. 46).

Scherthan, H. (2001). "A bouquet makes ends meet". *Nature reviews Molecular cell biology* 2.8, pp. 621–627 (cit. on p. 36).

Scherthan, H. et al. (2014). "Contrasting behavior of heterochromatic and euchromatic chromosome portions and pericentric genome separation in pre-bouquet spermatocytes of hybrid mice". *Chromosoma* 123, pp. 609–624 (cit. on p. 36).

Schleif, R. (1992). "DNA looping". *Annual review of biochemistry* 61.1, pp. 199–223 (cit. on p. 33).

Schmitt, A. D. et al. (2016). "Genome-wide mapping and analysis of chromosome architecture". *Nature reviews Molecular cell biology* 17.12, pp. 743–755 (cit. on p. 125).

Schoenfelder, S., T. Sexton, et al. (2010). "Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells". *Nature genetics* 42.1, pp. 53–61 (cit. on p. 33).

Schoenfelder, S., R. Sugar, et al. (2015). "Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome". *Nature genetics* 47.10, pp. 1179–1186 (cit. on p. 33).

Schücker, K. et al. (2015). "Elucidation of synaptonemal complex organization by super-resolution imaging with isotropic resolution". *Proceedings of the National Academy of Sciences* 112.7, pp. 2029–2033 (cit. on p. 18).

Schultz, D. C., K. Ayyanathan, et al. (2002). "SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins". *Genes & development* 16.8, pp. 919–932 (cit. on p. 113).

Schultz, D. C., J. R. Friedman, et al. (2001). "Targeting histone deacetylase complexes via KRAB-zinc finger proteins: the PHD and bromodomains of KAP-1 form a cooperative unit that recruits a novel isoform of the Mi-2$\alpha$ subunit of NuRD". *Genes & development* 15.4, pp. 428–443 (cit. on p. 113).

Scott, C. et al. (2018). "Proteomic profile of sex-sorted bull sperm evaluated by SWATH-MS analysis". *Animal reproduction science* 198, pp. 121–128 (cit. on p. 100).

Seabold, S. and J. Perktold (2010). "statsmodels: Econometric and statistical modeling with python". *9th Python in Science Conference* (cit. on p. 49).

Sertoli, E. (1865). "e lesistenza di particulari cellule ramificate nei canalicoli seminiferi dell'testicolo umno". *Morgagni* 7, p. 31 (cit. on p. 23).

Sexton, T. et al. (2012). "Three-dimensional folding and functional organization principles of the Drosophila genome". *Cell* 148.3, pp. 458–472 (cit. on p. 33).

Shami, A. N. et al. (2020). "Single-cell RNA sequencing of human, macaque, and mouse testes uncovers conserved and divergent features of mammalian spermatogenesis". *De-*

*velopmental cell* 54.4, pp. 529–547 (cit. on pp. 27, 28, 30, 31, 54, 64, 70, 72, 100, 135, 136).

Shettles, L. B. (1960). "Nuclear morphology of human spermatozoa". *Obstetrics & Gynecology* 16.1, pp. 10–14 (cit. on p. 100).

– (1961). "Human spermatozoa shape in relation to sex ratios". *Fertility and Sterility* 12.6, pp. 502–508 (cit. on p. 100).

Shi, J. et al. (2023). "Absence of MerTK disrupts spermatogenesis in an age-dependent manner". *Molecular and Cellular Endocrinology* 560, p. 111815 (cit. on pp. 93, 97).

Sin, H.-S. and S. H. Namekawa (2013). "The great escape: Active genes on inactive sex chromosomes and their evolutionary implications". *Epigenetics* 8.9, pp. 887–892 (cit. on p. 136).

Sleutels, F. et al. (2012). "The male germ cell gene regulator CTCFL is functionally different from CTCF and binds CTCF-like consensus sites in a nucleosome composition-dependent manner". *Epigenetics & chromatin* 5, pp. 1–21 (cit. on p. 130).

Sohni, A. et al. (2019). "The neonatal and adult human testis defined at the single-cell level". *Cell reports* 26.6, pp. 1501–1517 (cit. on pp. 27–29, 54, 62, 64, 70, 72, 73, 94–96, 100, 133, 135).

Song, W.-H. et al. (2018). "Effect of endocrine disruptors on the ratio of X and Y chromosome-bearing live spermatozoa". *Reproductive Toxicology* 82, pp. 10–17 (cit. on p. 100).

Soumillon, M. et al. (2013). "Cellular source and mechanisms of high transcriptome complexity in the mammalian testis". *Cell reports* 3.6, pp. 2179–2190 (cit. on pp. 54, 88, 97).

Spinnler, K. et al. (2010). "Glial cell line-derived neurotrophic factor is constitutively produced by human testicular peritubular cells and may contribute to the spermatogonial stem cell niche in man". *Human reproduction* 25.9, pp. 2181–2187 (cit. on p. 23).

Ståhl, P. L. et al. (2016). "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics". *Science* 353.6294, pp. 78–82 (cit. on p. 27).

Statello, L. et al. (2021). "Gene regulation by long non-coding RNAs and its biological functions". *Nature reviews Molecular cell biology* 22.2, pp. 96–118 (cit. on pp. 66, 71).

Steger, K. (1999). "Transcriptional and translational regulation of gene expression in haploid spermatids". *Anatomy and embryology* 199.6, pp. 471–487 (cit. on p. 71).

Street, K. et al. (2018). "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics". *BMC genomics* 19, pp. 1–16 (cit. on p. 84).

Stuart, T. et al. (2021). "Single-cell chromatin state analysis with Signac". *Nature methods* 18.11, pp. 1333–1341 (cit. on p. 41).

Svingen, T. and P. Koopman (2013). "Building the mammalian testis: origins, differentiation, and assembly of the component cell populations". *Genes & development* 27.22, pp. 2409–2426 (cit. on p. 31).

Takasaki, N. et al. (2014). "A heterozygous mutation of GALNTL5 affects male infertility with impairment of sperm motility". *Proceedings of the National Academy of Sciences* 111.3, pp. 1120–1125 (cit. on p. 77).

Talluri, T. R. et al. (2022). "Integrated multi-omics analyses reveals molecules governing sperm metabolism potentially influence bull fertility". *Scientific Reports* 12.1, p. 10692 (cit. on pp. 77, 94).

Tan, K. et al. (2020). "Transcriptome profiling reveals signaling conditions dictating human spermatogonia fate in vitro". *Proceedings of the National Academy of Sciences* 117.30, pp. 17832–17841 (cit. on pp. 28, 72).

Tang, W. W., S. Dietmann, et al. (2015). "A unique gene regulatory network resets the human germline epigenome for development". *Cell* 161.6, pp. 1453–1467 (cit. on p. 137).

Tang, W. W., T. Kobayashi, et al. (2016). "Specification and epigenetic programming of the human germ line". *Nature Reviews Genetics* 17.10, pp. 585–600 (cit. on p. 137).

Tian, G. G. et al. (2022). “Three-dimensional genome structure shapes the recombination landscape of chromatin features during female germline stem cell development”. *Clinical and Translational Medicine* 12.6, e927 (cit. on p. 130).

Turner, J. M. (2015). “Meiotic silencing in mammals”. *Annual review of genetics* 49, pp. 395–412 (cit. on p. 19).

Turner, J. M., S. K. Mahadevaiah, P. J. Ellis, et al. (2006). “Pachytene asynapsis drives meiotic sex chromosome inactivation and leads to substantial postmeiotic repression in spermatids”. *Developmental cell* 10.4, pp. 521–529 (cit. on p. 136).

Turner, J. M., S. K. Mahadevaiah, O. Fernandez-Capetillo, et al. (2005). “Silencing of unsynapsed meiotic chromosomes in the mouse”. *Nature genetics* 37.1, pp. 41–47 (cit. on p. 31).

Tüttelmann, F. et al. (2018). “Disorders of spermatogenesis: perspectives for novel genetic diagnostics after 20 years of unchanged routine”. *Medizinische Genetik* 30.1, pp. 12–20 (cit. on pp. 4, 31).

Umehara, T., I. Kawashima, et al. (2016). “Neuregulin 1 regulates proliferation of Leydig cells to support spermatogenesis and sexual behavior in adult mice”. *Endocrinology* 157.12, pp. 4899–4913 (cit. on p. 77).

Umehara, T., N. Tsujita, et al. (2019). “Activation of Toll-like receptor 7/8 encoded by the X chromosome alters sperm motility and provides a novel simple technology for sexing sperm”. *PLoS biology* 17.8, e3000398 (cit. on p. 100).

Van Alphen, M. et al. (1988). “Depletion of the spermatogonia from the seminiferous epithelium of the rhesus monkey after X irradiation”. *Radiation research* 113.3, pp. 473–486 (cit. on pp. 16, 72).

Vandereyken, K. et al. (2023). “Methods and applications for single-cell and spatial multi-omics”. *Nature Reviews Genetics*, pp. 1–22 (cit. on p. 71).

Vara, C. et al. (2019). “Three-dimensional genomic structure and cohesin occupancy correlate with transcriptional activity during spermatogenesis”. *Cell reports* 28.2, pp. 352–367 (cit. on pp. 36–38, 115, 126, 129, 130, 135, 137).

Vasiliauskaitė, L. et al. (2018). “Defective germline reprogramming rewires the spermatogonial transcriptome”. *Nature structural & molecular biology* 25.5, pp. 394–404 (cit. on p. 137).

Ventela, S. et al. (2003). “Intercellular organelle traffic through cytoplasmic bridges in early spermatids of the rat: mechanisms of haploid gene product sharing”. *Molecular biology of the cell* 14.7, pp. 2768–2780 (cit. on p. 113).

Vierstra, J. et al. (2020). “Global reference mapping of human transcription factor footprints”. *Nature* 583.7818, pp. 729–736 (cit. on pp. 46, 47).

Vries, M. de et al. (2012). “Human male meiotic sex chromosome inactivation”. *PloS one* 7.2, e31485 (cit. on p. 31).

Wang, C. and H. Liu (2022). “Factors influencing degradation kinetics of mRNAs and half-lives of microRNAs, circRNAs, lncRNAs in blood in vitro using quantitative PCR”. *Scientific Reports* 12.1, p. 7259 (cit. on p. 96).

Wang, M., X. Liu, et al. (2018). “Single-cell RNA sequencing analysis reveals sequential cell fate transition during human spermatogenesis”. *Cell stem cell* 23.4, pp. 599–614 (cit. on pp. 27, 31, 54, 64, 70, 100).

Wang, R.-S. et al. (2009). “Androgen receptor roles in spermatogenesis and fertility: lessons from testicular cell-specific androgen receptor knockout mice”. *Endocrine reviews* 30.2, pp. 119–132 (cit. on pp. 79, 92).

Wang, T. et al. (2019). “Essential role of histone replacement and modifications in male fertility”. *Frontiers in genetics* 10, p. 962 (cit. on p. 21).

Wang, X., J.-H. Choi, et al. (2013). “HOXC9 directly regulates distinct sets of genes to coordinate diverse cellular processes during neuronal differentiation”. *BMC genomics* 14.1, pp. 1–16 (cit. on p. 31).

Wang, Y., H. Wang, et al. (2019). “Reprogramming of meiotic chromatin architecture during spermatogenesis”. *Molecular cell* 73.3, pp. 547–561 (cit. on pp. 36–38, 115, 126, 127, 129, 130, 135, 137).

Wang, Z.-Y., E. Leushkin, et al. (2020). “Transcriptome and translatome co-evolution in mammals”. *Nature* 588.7839, pp. 642–647 (cit. on p. 30).

Wei, T. and V. Simko (2021). *R package 'corrplot': Visualization of a Correlation Matrix*. (Version 0.92) (cit. on p. 46).

Werner, A. et al. (2015). “Transpositional shuffling and quality control in male germ cells to enhance evolution of complex organisms”. *Annals of the New York Academy of Sciences* 1341.1, pp. 156–163 (cit. on p. 30).

Wu, X. et al. (2022). “Single-cell ATAC-Seq reveals cell type-specific transcriptional regulation and unique chromatin accessibility in human spermatogenesis”. *Human molecular genetics* 31.3, pp. 321–333 (cit. on pp. 33, 54, 73, 79, 93, 95).

Xia, B. et al. (2020). “Widespread transcriptional scanning in the testis modulates gene evolution rates”. *Cell* 180.2, pp. 248–262 (cit. on pp. 27, 30, 54, 64, 88, 100).

Yamaguchi, K. et al. (2018). “Re-evaluating the localization of sperm-retained histones revealed the modification-dependent accumulation in specific genome regions”. *Cell reports* 23.13, pp. 3920–3932 (cit. on p. 21).

Yan, F. et al. (2020). “From reads to insight: a hitchhiker's guide to ATAC-seq data analysis”. *Genome biology* 21, pp. 1–16 (cit. on p. 32).

Yanagimachi, R. (1994). *Mammalian fertilization. In. Knobil E, Neill J (editors). The Physiology of Reproduction* (cit. on p. 20).

Yang, E. et al. (2003). “Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes”. *Genome research* 13.8, pp. 1863–1872 (cit. on p. 96).

Yao, Y. and W. Dai (2012). “Shugoshins function as a guardian for chromosomal stability in nuclear division”. *Cell Cycle* 11.14, pp. 2631–2642 (cit. on p. 20).

Yin, Q. et al. (2022). “Revisiting chromatin packaging in mouse sperm”. *bioRxiv*, pp. 2022–12 (cit. on pp. 38, 39, 115, 130, 137).

You, Y.-A. et al. (2018). “2, 3, 7, 8-Tetrachlorodibenzo-p-dioxin can alter the sex ratio of embryos with decreased viability of Y spermatozoa in mice”. *Reproductive Toxicology* 77, pp. 130–136 (cit. on p. 100).

Young, M. D. and S. Behjati (2020). “SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data”. *Gigascience* 9.12, giaa151 (cit. on p. 42).

Yu, G., L.-G. Wang, Y. Han, et al. (2012). “clusterProfiler: an R package for comparing biological themes among gene clusters”. *Omics: a journal of integrative biology* 16.5, pp. 284–287 (cit. on p. 45).

Yu, G., L.-G. Wang, and Q.-Y. He (2015). “ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization”. *Bioinformatics* 31.14, pp. 2382–2383 (cit. on p. 45).

Zappia, L. and A. Oshlack (2018). “Clustering trees: a visualization for evaluating clusterings at multiple resolutions”. *Gigascience* 7.7, giy083 (cit. on p. 94).

Zhang, J., K. Eto, et al. (2011). “Neuregulins are essential for spermatogonial proliferation and meiotic initiation in neonatal mouse testis”. *Development* 138.15, pp. 3159–3168 (cit. on pp. 77, 94).

Zhang, Y., T. Liu, et al. (2008). “Model-based analysis of ChIP-Seq (MACS)”. *Genome biology* 9.9, pp. 1–9 (cit. on p. 42).

Zhao, L. et al. (2020). “Single-cell analysis of developing and azoospermia human testicles reveals central role of Sertoli cells”. *Nature communications* 11.1, p. 5683 (cit. on pp. 30–32, 54).

Zheng, S. C. et al. (2022). “Universal prediction of cell-cycle position using transfer learning”. *Genome biology* 23.1, pp. 1–27 (cit. on pp. 47, 80, 83).

Zhou, Y. et al. (2020). “Evaluation of single-cell cytokine secretion and cell-cell interactions with a hierarchical loading microwell chip”. *Cell reports* 31.4 (cit. on p. 98).

Zhu, Y.-T. et al. (2016). “Differences and similarities between extremely severe oligozoospermia and cryptozoospermia in intracytoplasmic sperm injection”. *Asian Journal of Andrology* 18.6, p. 904 (cit. on p. 31).

Zirkin, B. R. and V. Papadopoulos (2018). “Leydig cells: formation, function, and regulation”. *Biology of reproduction* 99.1, pp. 101–111 (cit. on p. 22).

# Appendix A

# Clinical parameters of samples processed in the multiome study

| | 1 | 2 | 3 |
|---|---|---|---|
| **Patient ID** | RG_CERA_012 SSC 3755 | RG_CERA_013 SSC 3759 | RG_CERA_042 SSC 3993 |
| **Age (Years)** | 32 | 33 | 41 |
| **Ejaculate Volume (ml)** | 4.6 | 0.7 | 4.7 |
| **Total Sperm Count (Millions)** | 0 | 0 | 0 |
| **Sperm Concentration (Million/ml)** | 0 | 0 | 0 |
| **FSH (normal range 1-7 U/l)** | 1.2 | 2.2 | 3.7 |
| **LH (normal range 2-10 U/l)** | 4.2 | 2 | 2.3 |
| **Total Testosterone (normal range >12 nmol/l)** | 34.5 | 24.9 | 18 |
| **SHBG (normal range 11-71 nmol/l)** | 51 | 55 | Not available |
| **Free Testosterone (normal range >250 pmol/l)** | 637 | 400 | Not available |
| **Prolactin (normal range <500 mU/l)** | 107 | 142 | Not available |
| **Estradiol (normal range <250 pmol/l)** | 88 | 82 | Not available |
| **DHT (normal range 0.5-2.0 nmol/l)** | 1.67 | 0.55 | Not available |
| **Testicular volume (ml)** | 30 | 22 | 19 |
| **Bergmann-Kliesch Score** | 7 | 8 | 8 |
| **% tubules with elongated spermatids** | 66 | 79 | 79 |
| **% tubules with round spermatids** | 1 | 0 | 0 |
| **% tubules with spermatocytes** | 26 | 12 | 13 |
| **% tubules with spermatogonia** | 6 | 2 | 1 |
| **% tubules sertoli cell only** | 1 | 0 | 2 |
| **% tubular shadows** | 0 | 7 | 5 |
| **Retrieval of sperm from the TESE** | Yes | Yes | Yes |
| **Karyotype** | 46,XY | 46,XY | Not available |
| **CFTR** | Not available | Heterozygous for F508del [c.1521_1523delCTT p.Phe508del], pathogen Poly(T)-Allele in Intron 8: T7/T9 | Not available |
| **AZF deletion** | No deletion | No deletion | Not available |

# Appendix B

# Antibody Dilutions

| Primary Antibody | | | | | |
|---|---|---|---|---|---|
| **Antibody** | **Company** | **Catalog Number** | **Host** | **Dilution for whole mount staining** | **Dilution for FACS** |
| spagnoli-mMAGEA4 | Prof.G.C. Spagnoli | - | Mouse | 1:20 | 1:20 |
| abcam-mMAGEA4 | Abcam | ab139297 | Mouse | 1:100 | 1:10 |
| UTF-1 | Millipore | MAB4337 | Mouse | 1:50 | 1:20 |
| DMRT1-Al647 | Santa Cruz | sc-377167 AF647 | Mouse | 1:100 | 1:200 |
| KIT | R&D | AF332 | Goat | 1:20 | 1:10 |
| KIT-APC | BD Pharmingen | 561118 | Mouse | 1:20 | 1:25 |
| KIT-Al647 | Biolegend | 313235 | Mouse | 1:40 | 1:6 |
| IgG (for spagnoli-mMAGEA4) | Biolegend | 400165 | Mouse | 1:20 | 1:20 |
| IgG (for abcam-mMAGEA4) | Abcam | ab18413 | Mouse | 1:100 | 1:10 |
| IgG (for UTF1) | Biolegend | 400165 | Mouse | 1:50 | 1:20 |
| IgG-Al647 (for DMRT1-Al647) | Biolegend | 400130 | Mouse | 1:100 | 1:200 |
| IgG (for KIT) | Sigma-Aldrich | 5256 | Goat | 1:20 | 1:10 |
| IgG-APC (for KIT-APC) | BD Pharmingen | 550854 | Mouse | 1:20 | 1:25 |
| IgG-Al647 (for KIT-Al647) | Biolegend | 400130 | Mouse | 1:40 | 1:6 |

| Secondary Antibody | | | | | |
|---|---|---|---|---|---|
| **Antibody** | **Company** | **Catalog Number** | **Host** | **Dilution for whole mount staining** | **Dilution for FACS** |
| Anti-Mouse IgG (H+L), Alexa Fluor 647 | Thermo Fisher | A-31571 | Donkey | 1:100 | - |
| Anti-Mouse IgG (H+L), Alexa Fluor 488 | Jackson Immunoresearch Laboratories | 715-546-150 | Donkey | 1:100 | - |
| Anti-Goat IgG H&L, Alexa Fluor 568 | Abcam | ab175474 | Donkey | 1:100 | - |

# Appendix C

# Journal Article (submitted)

# Gene regulatory control in human spermatogenesis and infertility

Elizabeth Ing-Simmons[1,2,3*], Jahnavi Bhaskaran[1,2,3*], Sara Di Persio[4], Jann-Frederik Cremers[5], Nadja Rotte[6], Sabine Kliesch[5], Corinna Friedrich[6], Sandra Laurentino[4], Nina Neuhaus[4], Juan M. Vaquerizas[1,2,3]

1. MRC London Institute of Medical Sciences, London, UK
2. Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London, UK
3. Max Planck Institute for Molecular Biomedicine, Münster, Germany
4. Centre of Reproductive Medicine and Andrology, Institute of Reproductive and Regenerative Biology, University of Münster, Münster, Germany
5. Department of Clinical and Surgical Andrology, Centre of Reproductive Medicine and Andrology, University Hospital of Münster, Münster, Germany.
6. Institute of Reproductive Genetics, University of Münster, Münster, Germany
*These authors contributed equally.

Correspondence should be addressed to JMV (j.vaquerizas@lms.mrc.ac.uk).

**Abstract**

Gametogenesis, the production of oocytes and sperm, ensures the faithful transmission of genetic material to the next generation. In males, spermatogenesis occurs in the testes, which contain the complete developmental progression of germ cells alongside the required supporting somatic cells. Defects in this fundamental process lead to infertility. However, the precise regulatory programs that govern spermatogenesis are unknown. Here, we generate a comprehensive gene regulatory network of human spermatogenesis by using single nucleus multiomics technology to carry out joint analysis of gene expression and chromatin accessibility at the single nucleus level. We identify stage-specific activity of both known and novel regulators of spermatogenesis and validate this using transcription factor footprinting. We demonstrate the predictive power of the network by linking predicted enhancers to genomic variants identified in cohorts of infertile men. By generating additional multiomics datasets from two infertile individuals, we identify transcription factors and target genes that are perturbed in infertility. This study provides a comprehensive single nucleus map of chromatin accessibility, gene expression, and gene regulation during normal and dysfunctional human spermatogenesis, which can be used to inform the interpretation of genes and genomic variation implicated in male infertility.

## Introduction

Gametogenesis, the production of haploid oocytes or sperm from diploid precursor cells, is a fundamental process that ensures the faithful transmission of genetic data to the next generation and therefore the continuation of a species. Defects in this process can lead to infertility. In humans, infertility is estimated to affect one in six people worldwide during their lifetime, with around half of those cases involving a male factor[1,2]. However, only around 30% of infertile men receive a causal diagnosis[3]. This limits the ability of clinicians to provide counselling about potential treatments, transmission of infertility to offspring, and potential health impacts. In addition, while assisted reproduction technologies can allow conception, these are associated with a lower rate of pregnancy[4] and poorer health outcomes in offspring[5], indicating an incomplete understanding of the sperm characteristics required for successful reproduction.

Sperm is produced within the testis, a highly complex tissue from a morphological perspective as well as from a cell composition and dynamics point of view. In healthy adult males, the testis contains all of the morphological structures and cell types required to drive spermatogenesis[6]. Spermatogenesis comprises a continuum of proliferation and differentiation states: from spermatogonia, including spermatogonial stem cells, through spermatocytes which undergo meiosis, to haploid spermatids, which undergo spermiogenesis, the process of maturation into spermatozoa (Fig 1A). These germ cells reside within seminiferous tubules alongside Sertoli cells, which provide physical and nutritional support to them[7]. The tubules are surrounded by peritubular myoid cells (PTMs), and the interstitial compartment between tubules includes testosterone-producing Leydig cells, immune cells, and vasculature-associated pericytes and endothelial cells[8].

In accordance with this diverse cell composition, previous work has shown that the testis has one of the most complex transcriptomes in the human body[9,10]. Recent single cell (sc) and single nucleus (sn) approaches have allowed researchers to examine this tissue complexity and characterise the heterogeneity of cell types and states within the human testis[11], as well as the origins of these cell types during embryonic development[12] and puberty[13], and their molecular evolution across species[14].

Despite these advances, an understanding of the molecular mechanisms that regulate spermatogenesis is lacking. The chromatin state of germ cells undergoes dramatic changes during spermatogenesis, including transcriptional silencing of the sex chromosomes during meiosis[15], and large-scale remodelling during spermiogenesis to facilitate the replacement of histones by protamines[16], which allow for tight packaging of the genome inside the compact sperm head. In addition, scATAC-seq identified changes in transcription factor motif usage across spermatogenesis[17]. However, a comprehensive understanding of the key transcription factors and regulatory elements which drive normal human spermatogenesis requires joint single-cell assessment of both chromatin state and gene expression in order to robustly infer their relationships and regulatory dynamics. Here, we produce a single-nucleus atlas of chromatin accessibility and gene expression in human testis, and combine this data with transcription factor motif enrichment to define a gene regulatory network that drives developmental transitions during spermatogenesis. We use this network to identify targets of

infertility-associated single-nucleotide polymorphisms (SNPs). Finally, we produce single nucleus multiome data in infertile men and identify perturbations of the gene regulatory network in these individuals, including of key infertility-associated genes.

## Results

### Single nucleus multiomics captures the complexity of the testis

We generated single nucleus multiome (snRNA-seq and snATAC-seq) data from testis biopsies from three men with normal sperm development according to clinical parameters (Supplementary Table 1). After standard quality control and filtering, we retained > 19,000 high quality cells for downstream analysis, with a mean of ~4,700 RNA molecules (unique molecular identifiers) and a mean of ~2,250 ATAC-seq peaks detected per cell. RNA and ATAC-seq data were separately processed and then integrated using a weighted nearest-neighbour (WNN) approach[18]. Based on the expression of known marker genes for different testis cell types[11] (Fig S1), we identified 14 germline clusters and seven somatic clusters (Fig 1B). Four additional clusters could not be assigned to a known cell type. These clusters showed mixed germline-somatic expression (Fig S1) and/or strongly biased RNA/ATAC contributions (Fig S2), and were therefore excluded from further analysis. Importantly, the chromatin accessibility data allowed us to further resolve cluster identities that were unclear based on expression data alone. Cluster 15 has low expression of most marker genes (Fig S1). However, we observed that this cluster exhibits a bimodal distribution of ATAC-seq reads from chromosome X, indicating the separation of cells into chromosome X- or Y-bearing cells after meiosis I (Fig S3A). Along with the expression of *C9orf116*, a previously identified marker of secondary spermatocytes[19], we therefore consider cluster 15 to represent the rare population of cells undergoing meiosis II and not yet expressing spermatid markers. Among the somatic cells we could identify distinct clusters representing Sertoli cells, PTMs, endothelial cells, pericytes, and testicular immune cells. In comparison with previous scRNA-seq datasets, we identified a lower proportion of spermatids and we do not identify a distinct cluster of Leydig cells [20,21]. However, we identify a larger proportion of spermatogonia (Fig S4). This is likely due to differences in nuclei isolation efficiency across different cell types.

Trajectory inference using Slingshot[22] recapitulates the known continuum of spermatogenesis with a main trajectory that includes more than 99% of germline cells (inset, Fig 1B). Differential gene expression analysis identified a total of ~12,000 genes displaying cluster-specific expression, including many known markers for the different cell types in the testis (Fig 1C). Examining chromatin accessibility at these loci shows dynamic promoter and gene body accessibility which correlates with gene expression (Fig 1D and Fig S5). In addition, we can also identify distal peaks with stage-specific accessibility during spermatogenesis, which represent putative enhancers. An example is shown at the *TNP1* locus, where a peak to the left of the gene gains accessibility in early spermatids.

To understand how changes in chromatin accessibility might be regulated, we examined the chromatin accessibility of almost 700 transcription factor motifs[23], revealing clear cluster-specific patterns of motif activity that could be used to identify regulators of key transitions

during spermatogenesis (Fig 1E). Our chromatin accessibility data also captures the silencing of unsynapsed regions of the sex chromosomes during meiosis (meiotic sex chromosome inactivation; MSCI), shown by the global decrease in signal on the X chromosome in pachytene/diplotene (Fig S3). In addition, there is a marked drop in the fraction of ATAC-seq reads falling in discrete peaks after meiosis (Fig S3B). This broader chromatin opening may be linked to widespread expression in these stages[9,10] and may also reflect changes in chromatin state that precede replacement of histones by protamines[16]. Altogether, our single nucleus multiome atlas captures the gene expression and regulatory dynamics of a comprehensive repertoire of germ cells throughout spermatogenesis, and of their supporting somatic cells.

**Construction of a testis gene regulatory network**

To further characterize the gene regulatory landscape driving human spermatogenesis, we sought to predict enhancer elements, their candidate upstream regulatory transcription factors (TFs), and their target genes. To do so, we used SCENIC+ to combine the chromatin accessibility and gene expression data with TF motif enrichment[24]. SCENIC+ uses this data to identify, for every input TF, its predicted target regions and its predicted target genes; each set of a TF, predicted enhancers, and targets is called an eRegulon (Fig 2A). We identify 62 high confidence eRegulons with strong correlations between TF expression, predicted enhancer accessibility and target gene expression. These include 50 activating eRegulons targeting an average of 1,250 regions and 620 genes, and 12 repressive eRegulons targeting an average of 300 regions and 150 genes. In total, 56 TFs regulate over 10,000 unique genes and 23,000 accessible regions of which ~12,000 are distal predicted enhancers. We clustered these eRegulons into 4 groups based on target gene activity, identifying eRegulons that are active primarily in undifferentiated spermatogonia (Group 1, 16 eRegulons), differentiating spermatogonia and/or spermatocytes (Group 2, 12 eRegulons), spermatids (Group 3, 8 eRegulons) or somatic cells (Group 4, 28 eRegulons) (Fig 2B, Fig S6). Groups 2, 3, and 4 include both positive and negative regulators of target gene expression, while Group 1 contains only positive regulators. Known regulators of specific cell types show the expected patterns. For example, DMRT1, a key regulator of sex determination and spermatogenesis that is highly expressed in differentiating spermatogonia[25–27], and CREM, a TF involved in maturation of spermatids [28,29], are included in Group 2 and Group 3 respectively (Fig S6).

We next combined these 62 high confidence eRegulons to generate a gene regulatory network of the human testis (Fig 2C). Notably, TFs active in somatic cells or spermatogonia tend to reside in the dense central part of the network, while the periphery of the network includes genes and TFs active in spermatids. This tendency can be quantified by measuring betweenness centrality, representing the frequency with which a given TF lies on regulatory pathways that link other pairs of genes. TFs with higher betweenness centrality thus have more influence over other parts of the network. In line with the visual observations, TFs active in spermatids (Group 3) have significantly lower betweenness centrality than those active in undifferentiated spermatogonia (Group 1) and somatic cells (Group 4) (Fig 2D). Examining the number of regulators (in degree) and number of target genes (out degree) for these TFs indicates that this is at least partially driven by spermatid-specific TFs having a

lower in degree, while TFs active in spermatogonia have high numbers of both target genes and regulators, indicating extensive cross-regulation between TFs active at this stage.

As an illustrative example of gene regulation predicted by SCENIC+, we focus on the *RNF17* locus (Fig 2E). RNF17 is a RING finger domain protein that is involved in piRNA regulation and is essential for spermiogenesis in mouse [30,31]. While the promoter of *RNF17* shows accessibility throughout the germline, the gene has cluster-specific expression, with the strongest expression in zygotene and pachytene/diplotene, followed by spermatogonia (particularly State 0B and State 1). We identify two candidate enhancers for *RNF17* (red bars within grey highlights in Fig 2E), one of which is active in undifferentiated spermatogonia (left) and the other that is primarily active in zygotene and pachytene/diplotene (right). The first is associated with positive regulation by LIN28B, and the second is positively regulated by LIN28B and CLOCK, and repressed by JUN. To support the activity of these activating TFs in the stages where the enhancers are active, we carried out transcription factor footprinting analysis using TOBIAS[32]. Results for representative motifs bound by LIN28B (top) and CLOCK (bottom) are shown in Fig 2F, supporting that the strongest activity of these TFs is in pachytene/diplotene.

Additional examples are shown in Fig S7. Notably, an enhancer of the E3 ubiquitin-protein transferase *RMND5A* that is predicted by SCENIC+ to be regulated by DMRT1 overlaps with ChIP-seq data for DMRT1 binding[33] (Fig S7A). Footprinting analysis for DMRT1 confirms its expected activity in differentiating spermatogonia, coincident with the accessibility of this predicted enhancer and the expression of *RMND5A* (Fig S7C). Assessing genome-wide overlap of DMRT1 ChIP-seq peaks and enhancers predicted to be regulated by DMRT1 shows a highly significant overlap (27.6% compared to 2.6% of all predicted enhancers, permutation test $p = 0.001$), despite the fact that this data is derived from whole testis and therefore lacks power to detect weaker or highly cell type specific DMRT1 binding sites. For *VGLL3*, a cofactor for TEAD transcription factors[34,35] which is expressed in both spermatids and PTMs, SCENIC+ identifies two putative regulatory regions (Fig S7B), one of which is predicted to regulate *VGLL3* in spermatids via the activity of the TFs ATF5 and CREM, and the other of which is predicted to be activated in PTMs by EBF1 and NR2F2.

Together, these examples illustrate that we can predict cell type specific enhancers, for both germline and somatic cells, regulated by novel as well as known regulators of key spermatogenesis transitions. The activity of the assigned TFs is supported by TF footprinting and independent ChIP-seq data. The full set of high-confidence TF-target gene pairs is given in Supplementary Table S2. Overall, our multiome data allows us to identify a robust network of cell type specific gene regulation for human spermatogenesis.

## The spermatogenesis gene regulatory network identifies targets of infertility-associated SNPs

Next, we explored the association of these eRegulons with male infertility phenotypes. To date, three GWAS studies of male infertility have identified a total of 7 SNPs directly associated with non-obstructive azoospermia – defined as the absence of sperm in the ejaculate due to impaired spermatogenesis – in either Han Chinese or European populations

(Supplementary Table S3)[36–38]. As is common in GWAS results[39–41], most (6/7) of these SNPs fall in non-coding regions and their mechanistic link with infertility is unclear. Therefore, we assessed whether our gene regulatory network could be used to assign target genes that could explain their link with infertility. Since none of these 6 non-coding SNPs directly overlapped a SCENIC+ predicted enhancer, we extended our analysis to 'proxy' SNPs in high linkage disequilibirum (LD; $R^2 > 0.8$) with each candidate SNP in the matching study population (Han Chinese or European), using LD data from the 1000 Genomes Project Phase 3[42,43]. This approach identified between 3 and 147 proxy SNPs within 500kb of each of the original SNPs. For one of the original SNPs, rs2477686, we were able to identify two SCENIC+ predicted enhancers overlapping its proxy SNPs. These enhancers lie within the gene body of *PLCH2* and are predicted to regulate the nearby genes *MMEL1* (Fig 3A-B) and *TNFRSF14* (Fig S8). Fig 3B shows the enhancer predicted to regulate *MMEL1*, a membrane metallo-endopeptidase which has been previously implicated in male fertility. Knockout of its mouse homolog results in reduced fertility, while a *Drosophila* ortholog has also been found to be required for proper sperm function[44,45]. This enhancer has increased accessibility in spermatids and is predicted to be regulated by KLF17, a transcription factor previously identified as a tumour suppressor and regulator of pluripotency in human embryonic stem cells [46,47]. Consistent with the enhancer accessibility, the activity of the KLF17 eRegulon at the level of both target genes and regions, as well as the expression of both *KLF17* and *MMEL1*, is increased in spermatids (Fig 3C-D). The second enhancer (Fig S8) is predicted to regulate both *TNFRSF14* and *MMEL1*. This enhancer has increased accessibility in PTMs and pericytes and is predicted to be a target of 5 TFs regulating *TNFRSF14*, including EBF1 and NFIC with PTM-enriched eRegulon activity (Fig S8). In conclusion, our gene regulatory network allows us to link cell type specific gene regulatory modules to SNPs associated with infertility. Therefore, this network will provide a valuable resource for the interpretation of future GWAS carried out for fertility-related phenotypes.

**Gene regulatory network perturbation in infertility**

Given the current limitations of *in vitro* spermatogenesis approaches[48], and therefore the inability to perform perturbation studies to assess the validity of our gene regulatory network in human samples, we sought an alternative approach to evaluate its biological relevance. To do so, we performed single nucleus multiomics experiments on two individuals with cryptozoospermia (sperm concentration below 0.1 million/ml[3]). We identified a total of >17,000 high quality cells, with a mean of ~3,900 RNA molecules (UMIs) and a mean of ~2,400 ATAC-seq peaks detected per cell. After integration with the normal spermatogenesis samples, we identified all the expected germline and somatic cell types, including Leydig cells (Fig 4A, Fig S9). Visualisation of the data highlighted a dramatic depletion of spermatids in the cryptozoospermia samples, which is supported by quantification of the fraction of cells in each cluster originating from each condition (Fig 4B). Despite sharing the same clinical phenotype of cryptozoospermia, differential gene expression and chromatin accessibility analysis revealed distinct molecular phenotypes in the two samples, with a much greater level of differential gene expression and accessibility in sample 2 (Fig 4C, Fig S10). Exome sequencing analysis[49] revealed that the first individual sampled carries a mutation in *M1AP*, a gene implicated in regulation of meiotic recombination[49–51]. Accordingly, differences in gene expression and accessibility in this

sample were minimal and largely restricted to pachytene and later stages. Therefore, we focused our analysis on Crypto 2, which shows widespread downregulation of genes from spermatogonia onward and increases in accessibility during meiosis (Fig 4C, Fig S11). Amongst the differentially expressed genes are several TFs from our high-confidence gene regulatory network (Fig 4D). We therefore assessed whether differential expression of TFs is associated with differential expression of their target genes. Across all combinations of clusters and TFs, in 46.6% of analyses there was a significant association of target genes with the same direction of change in expression as their positive regulatory TF (Fisher's exact test, Benjamini-Hochberg adjusted p-value < 0.05). To visualise these significant associations, we measured the average log2FC of expression of these target genes, subtracting the average log2FC of all genes in the cluster to control for global shifts in expression (Fig 4E), and plotted this against the log2FC of the TFs expression (Fig 4F). This reveals a clear correlation in the direction and magnitude of expression of target genes and their corresponding TF and provides evidence for the predictive power of the TF-target associations in ourgene regulatory network and support for their biological relevance. Finally, we assessed whether genes with known roles in spermatogenesis are differentially expressed in this individual and could explain the infertility phenotype. Out of 156 differentially expressed genes in Crypto 2, across all clusters, 17 are required for male fertility in mice, according to knockouts performed by the International Mouse Phenotyping Consortium[52] (Fisher's Exact Test p = 0.003). These include *HSP90AA1*, a molecular chaperone that has been implicated in regulation of piRNAs[53], *SPINK2*, a serine protease inhibitor that is required for acrosomal biogenesis and has been implicated in infertility in humans as well as mice[54], and *CCDC34*, a flagellar protein thought to be required for proper flagella formation in both mouse and human sperm [55]. In addition, manual examination of downregulated genes identified additional genes with known functions in fertility (Fig 4G), such as *DAZL*, an RNA binding protein that is essential for both male and female gametogenesis[56], with roles in the specification of germ cells[57] as well as in spermatogonia and spermatocytes [58]. Other examples include *HORMAD1*, which is associated with the synaptonemal complex that is essential for both male and female gametogenesis in mice[59], and has recently been implicated in male infertility in humans[60], and *FSIP2*, mutations in which are associated with defects in sperm flagellum and acrosome formation in human and mouse [61–63]. Overall, our analysis identifies multiple misregulated genes in cryptozoospermia which may play functional roles in the development of the clinical phenotype.

## Discussion

Here we comprehensively characterise the dynamics of chromatin accessibility and gene expression throughout normal human spermatogenesis and assess how they are altered in two individuals with cryptozoospermia. Although previous studies had separately analysed scRNA-seq and scATAC-seq data from human testis, without paired data from the same single cells the relationships between these modalities could only be inferred. Notably, we identify genes where promoter and gene body accessibility correlates well with gene expression (Fig 1D) as well as genes with promoter accessibility throughout spermatogenesis despite cluster-specific expression (Fig 2E, S8A). This underscores the limitations of inferring gene expression from accessibility and highlights the advantages of multiome data from the same nuclei. Obtaining such single nucleus multiome data allowed

us to construct a robust gene regulatory network and identify 62 high confidence regulatory modules active during different stages of spermatogenesis and in the supporting somatic cell types of the testis. These regulatory modules are driven by transcription factors with known roles in spermatogenesis, such as DMRT1, and by novel candidate regulators, such as KLF17. Together, this data sheds light on how and when known regulators of spermatogenesis act as well as providing further candidates for future study of the molecular mechanisms controlling male fertility.

The gene regulatory network analysis revealed that regulators active in spermatids are more peripheral to the network, with lower betweenness centrality than regulators active in spermatogonia and somatic cells. Previous work has identified correlations between network centrality and evolutionary rates of human genes, with younger genes and those undergoing faster turnover having lower centrality[64,65]. The peripheral location of spermatid TFs is therefore consistent with recent work showing that genes expressed in late spermatogenesis are younger, and evolving faster, than those expressed earlier[14]. This network topology may reflect a more recent evolutionary origin of these TFs and their target genes, or it may also provide a framework for rapid evolutionary changes to occur without disruption of the conserved core of the network[66,67].

Apart from a handful of well characterised loci[50,68], there is surprisingly little information about genomic variation associated with male infertility. In particular, the difficulty in identifying developmental stage-specific regulatory elements from heterogeneous tissue containing all stages of spermatogenesis has limited the ability to interpret functionality of non-coding variants. Here we have shown that the regulatory interactions present in our gene regulatory network can be used to identify potential functional links between GWAS SNPs for infertility and candidate genes. This provides a crucial resource for future GWAS studies, as well as studies utilising whole-genome or exome sequencing of infertile men.

Furthermore, by carrying out sn multiome experiments in infertile men, we identify regulatory modules which are perturbed in infertility, including genes that are known to have key roles in spermatogenesis. Notably, we find very different molecular phenotypes of the two infertile individuals studied here. This highlights the distinct molecular mechanisms that can underlie male infertility and underscores the importance of obtaining multiple biological replicates and assessing their similarities.

In summary, here we provide a comprehensive overview of chromatin accessibility and gene expression at the single cell level in the human testis, and demonstrate how to use this data to construct a gene regulatory network. This unique resource will be instrumental in aiding the interpretation of clinical studies and GWAS of male infertility, which will lead to an increased number of causal diagnoses for affected individuals. Ultimately, we envisage that this data will aid in the development of *in vitro* gametogenesis approaches and personalised medicine treatments for male infertility.

# Methods

## Selection of human testicular biopsies

We obtained testicular biopsies of patients from the Centre of Reproductive Medicine and Andrology (CeRA) in Münster, Germany. Normal spermatogenesis samples were obtained from patients who had been diagnosed with obstructive azoospermia due to a previous vasectomy, a congenital malformation or physical obstruction of the vas deferens, but had normal testicular volume, normal FSH levels and qualitatively and quantitatively normal spermatogenesis. Cryptozoospermic patients had a sperm concentration < 0.1 million/ml (cryptozoospermia) in the ejaculate, a reduced testicular size, and, in most cases, elevated FSH levels due to a severe impairment of spermatogenesis itself. All patients had spermatozoa in their TESE samples, regardless of histological results. A complete list of the measured clinical parameters can be found in Supplementary Tables S1 and S4 for normal and cryptozoomspermic patients respectively. When undergoing surgery for microdissection testicular sperm extraction (mTESE) or histological evaluation at the Department of Clinical and Surgical Andrology (University Hospital in Münster, Germany), one additional testicular sample was excised for this study after written informed consent. Ethical approval was obtained for this study (Ethics Committee of the Medical Association of Westphalia-Lippe and University of Münster 2010-578-f-S).

## Preparation of single cell suspensions from human adult testis biopsy

A two-step enzymatic digestion was performed as detailed in [69–71]. Briefly, the tissue sample was placed into 6 cm culture dish with MEM-alpha media (Life Technologies GmbH, Gibco, Darmstadt, Germany) and dissected into fragments of 1mm or less using a blade. It was then transferred into a 15 ml tube and centrifuged for 5 minutes at 1500 rpm and 10°C. The cell pellet was resuspended in MEM-alpha media containing 1mg/ml filter sterilized collagenase IA (C9891, Sigma) and incubated at 37°C until tubules separated. The tube was inverted every 5 minutes. The reaction was stopped by adding 5 ml MEM-alpha media with 10% FBS and 1% Pen/Strep. It was centrifuged at 10°C at 1500 rpm for 4 minutes, the supernatant discarded and incubated with 2ml filter sterilised Hank's balanced salt solution (Life Technologies GmbH, Gibco) containing 2 mg Trypsin (27250-018, Gibco) and 11 mg DNase I (DN25, Sigma) for 7-10 minutes at 37°C. The tube was inverted every 3 minutes. This was followed by strong pipetting to obtain a single-cell suspension. The reaction was stopped as outlined above and cells were washed twice with MEM-alpha. Then, erythrocytes were removed by incubating the cells with haemolysis buffer (0.83% NH4Cl solution) for 3 minutes at room temperature. Enzymatic digestion was stopped as outlined above. The cells suspension was filtered through a 70 µm strainer to remove cell clumps and debris. It was pelleted, washed and the cells were resuspended in 2 ml medium prior to cell counting using the trypan blue exclusion method.

## Generation of paired scATAC-seq and scRNA-seq libraries

200,000-400,000 cells were centrifuged at 500g for 10 minutes at 4°C. The supernatant was discarded, and the pellet resuspended in 50 µl Wash Buffer (20 mM HEPES pH 7.5, 150 mM

NaCl, 0.5 mM Spermidine, 1X Protease Inhibitor Cocktail). It was centrifuged at 500g for 10 minutes at 4°C and supernatant was discarded. The pellet was incubated with 45 µl Lysis Buffer (wash buffer with 0.01% NP40, 0.01% Digitonin, 1mM DTT and 1 U/µl RNAse inhibitor (3335399001, Sigma)) for 25-30 seconds on ice. The reaction was stopped with 50 µl Neutralising Buffer (lysis buffer with 2mM EDTA). It was again centrifuged at 500g for 10 minutes at 4°C, the supernatant discarded, and the pellet washed with 45 µl 1X Dilute Nuclei Buffer (20X Nuclei Buffer provided with 10x scATAC + GEX multiome kit). The pellet was resuspended in 10 µl 1X Dilute Nuclei Buffer solution and 2 µl sample was taken for counting with Trypan Blue.

9000-10,000 nuclei were loaded for the 10x scATAC + GEX multiome transposase reaction and the library prepared according to manufacturer's instructions in Chromium Next GEM Single Cell Multiome ATAC + Gene Expression Reagent Kits User Guide (Revision A). The libraries were sequenced in NovaSeq 6000 with paired end sequencing.

RNA-seq libraries were sequenced to a depth of 40,000 read pairs/nucleus and ATAC-seq libraries to a depth of 50,000 read pairs/nucleus according to 10x sequencing guidelines.

**scATAC-seq and scRNA-seq analysis**

Read mapping and quantification

The fastq files of individual samples were processed with cellranger-arc v2.0.0 and mapped to refdata-cellranger-arc-GRCh38-2020-A-2.0.0.tar.gz. For each sample, cellranger-arc count was used with default parameters to generate a filtered peaks x cells counts matrix and a filtered genes x cells counts matrix. The filtered matrices of each sample were loaded into R and further processed with Seurat v4.1.1[18] and Signac v1.7.0[72].

RNA library QC

Ambient RNA contamination was estimated and removed using soupX v1.6.1 [73]. Next, doublets were removed using scDblFinder v1.10.0 [74], using a doublet rate of 0.077 or 0.069, corresponding to the multiplet rate for 10,000 and 9,000 recovered cells respectively in the 10x multiome scATAC + GEX user guide (Revision A). Cells with > 20% mitochondrial reads, fewer than 500 UMIs or fewer than 250 detected genes were removed. These cut-offs were determined using the isOutlier() function from scater v1.24.0 [75] or by visually examining the distribution with the aim of retaining as many high quality cells as possible.

ATAC library QC

Doublets were removed with using scDblFinder v1.10.0 [74], using a doublet rate of 0.077 and 0.069 for 10,000 and 9,000 recovered cells respectively, and aggregateFeatures = TRUE. Cells with fewer than 150 detected peaks or fewer than 250 fragments were removed. These cut-offs were chosen as described above. Cells with FRiP < 0.20 or TSS enrichment < 2 were removed.

Peak calling

Peaks were called on the remaining high quality cells grouped by Cellranger-identified cluster, using macs2 [76] via the Seurat function CallPeaks(). Peaks on non-standard chromosomes and those in genomic blacklist regions were removed. A new assay was created by quantifying counts in each of the remaining peaks with function FeatureMatrix() from Seurat taken for all downstream processing.

**Normalization, sample integration, dimensional reduction**

The RNA and ATAC-seq assays of the biological replicates were separately integrated using Seurat.

RNA-seq assay

Genes x cells matrices were normalised with scran v1.24.0[77] as follows. The Seurat object was converted to a SingleCellExperiment (SCE) object and clustered using the quickCluster() function. Size factors were calculated using computeSumFactors(). The logNormCounts() function from scuttle v1.6.2[75] was used to calculate normalised counts with log = FALSE. The normalised counts were then log-transformed with the log1p() function and added to the data slot of the Seurat object.

The top 4000 variable features were identified with Seurat's FindVariableFeatures() function with selection.method = "vst". The datasets were then integrated using the Seurat functions SelectIntegrationFeatures(), FindIntegrationAnchors(), and IntegrateData(). Finally, the combined integrated matrix was then scaled and centred with function ScaleData() from Seurat and PCA was performed using RunPCA().

ATAC-seq assay

The union of peaks called in each individual snATAC-seq dataset was taken and all datasets were re-quantified using this unified peak set. The bottom 5% of features were removed and TF-IDF and SVD were carried out to produce an LSI matrix. Datasets were then merged and the LSI calculation repeated. The LSI embeddings were integrated using the functions FindIntegrationAnchors() (using all peaks and reciprocal LSI) and IntegrateEmbeddings().

Weighted nearest neighbour (WNN) identification, clustering, and cluster annotation

The Seurat function FindMultiModalNeighbors() was used on the integrated RNA data (PCA dimensions 1:20) and ATAC data (LSI dimensions 2:20) with k.nn = 25 for normal samples and k.nn = 30 when combining normal and crypto samples. Clustering was carried out on the resulting WNN network using the FindClusters() function with res = 0.8 for normal samples and res = 0.9 when combining normal and crypto samples. Clusters were annotated based on the expression of key marker genes [11,21].

**scRNAseq analysis for scRNA vs snRNA comparison**

Read mapping

The fastq files of individual samples were processed with cellranger v6.0.1 and mapped to refdata-gex-GRCh38-2020-A.tar.gz. For each sample, cellranger count was used with default parameters to generate a filtered genes x cells counts matrix.

Quality control filtering

The filtered matrices of each sample were loaded into R and further processed with Seurat v4.0.3. Doublets were removed using scDblFinder() v1.10.0 with the following doublet rates: 0.008 for Guo et al. 2018 [20], 0.046 for Di Persio and Tekath et al. 2021 [21], and 0.077 or 0.069 for 10,000 or 9,000 cells recovered in this study, according to the respective 10x user guides. Cells with > 24% mitochondrial reads, nFeature_RNA < 350 (< 1000 for this study) and > 9500 were removed. These cut-offs were based on function isOutlier() from scater v1.20.1 or by visually examining the distribution with the aim of retaining as many high quality cells as possible. Sample N1 from Di Persio and Tekath et al. 2021 was excluded from the analysis due to high ambient RNA contamination that resulted in higher than estimated number of droplets being called as cells by Cellranger.

Each dataset was normalised with scran v1.24.0 as described above for multiome data. The top 4000 variable features were identified with Seurat function FindVariableFeatures() with selection.method = "vst".

Normalization, sample integration, dimensional reduction

Datasets were then integrated with Seurat using the Seurat functions SelectIntegrationFeatures(), FindIntegrationAnchors(), and IntegrateData(), with normalization.method = "LogNormalize". The integrated assay was scaled and centred with Seurat function ScaleData() with all genes as features and PCA was performed using RunPCA().

Clustering and cluster annotation

Cells were clustered using Seurat function FindClusters() on the integrated RNA with dims 1:30 and resolution = 0.5. Clusters were annotated based on key marker gene expression, as used in Di Persio & Tekath et al. 2021[21].

**Trajectory inference**

Trajectory inference was carried out using Slingshot v2.4.0 [22], using the first 20 dimensions of the PCA reduction of the RNA data after ambient RNA removal. Only germline clusters were included. The cluster corresponding to Spermatogonia State 0 was specified as a trajectory start point. This produced two lineages. The main lineage shown in Fig 1B (inset) includes > 99% of germline cells. A secondary lineage only includes clusters representing undifferentiated spermatogonia (37% of germline cells, not shown).

## Differential gene expression

Identification of cluster-specific enriched genes was carried out using the FindAllMarkers() function from Seurat, using method "MAST" [78] with biological replicate as a latent variable. Only positively enriched markers with an adjusted p value < 0.05 were kept.

Identification of differentially expressed genes between normal and cryptozoospermic samples was carried out for each cluster independently using Seurat FindMarkers() with method "wilcox". Genes with an adjusted p value < 0.05 and an absolute log2 foldchange in expression of > 1 were kept.

## GO term enrichment

The top 100 most significantly enriched genes per cluster were used as input for GO term enrichment analysis with clusterProfiler (v4.7.1.003) function enrichGO [79], using all genes included in the differential gene expression analysis as the background gene set. The results were simplified to remove terms with high semantic similarity using method = "Wang" and cutoff = "0.7". For visualisation, the most significantly enriched (based on BH-adjusted p value) term per cluster was selected and enrichment of these terms across all clusters plotted. For GO term enrichment in cluster-specific differentially expressed genes between normal and Crypto samples, all genes with BH-adjusted p value < 0.05 and an absolute log2 foldchange in expression of > 1 were selected. Up- and downregulated genes were analysed separately, as described above.

## Differential accessibility

Identification of differentially accessible peaks between normal and cryptozoospermic samples was carried out for each cluster independently using Seurat FindMarkers() with method "LR". The total number of reads in peaks was used as a latent variable and only peaks with signal in at least 5% of cells in the cluster were assessed. Peaks with an adjusted p value < 0.05 were kept.

## Motif activity

Motif activity was calculated using the Arch v1.0.2 R[80] implementation of ChromVar[81]. Cellranger output filtered matrices were pre-processed using ArchR, keeping cells with a minimum of 80 fragments and a minimum TSS enrichment value of 1.45. Cells were then further filtered to keep only those that had passed the filtering and quality control procedures described above. Consensus motifs from Vierstra et al. 2020 [23] obtained from https://resources.altius.org/~jvierstra/projects/motif-clustering-v2.0beta/ were used for motif activity calculations using ArchR with default parameters. The resulting ChromVar Z-scores were then added back to the integrated Seurat object for downstream analysis.

## SCENIC+

ATAC-seq counts and cell metadata were exported to text files from the integrated Seurat object and used to construct a cisTopic object using pycisTopic v1.0.2.dev8+g848f78b. Topic modelling was carried out using pycisTopic with 500 iterations, alpha = 50, eta = 0.1, and numbers of topics between 5 and 40. The optimal number of topics was determined to be 30 based on the suggested metrics. This selected set of topics was then binarized in two ways: using the Otsu method and by taking the top 3000 regions per topic. In addition, differentially accessible regions across clusters were identified using default parameters. These sets of regions were then used as input into pycisTarget (v1.0.2.dev8+g48af509). pycisTarget was used to calculate motif enrichment using the provided motif rankings and scores for motifs from the 2022 SCENIC+ motif collection (database version "hg38_screen_v10_clust") [24].

The cisTopic object and motif enrichment results were combined with and RNA-seq counts after ambient RNA removal exported from Seurat to construct a SCENIC+ object [24]. SCENIC+ (v0.1.dev446+g5cf2469) was then run with default parameters. Search space for putative regulatory regions was 150kb upstream and downstream of each gene. Standard filtering was applied and returned 521 eRegulons, with a total of 471 TFs and 13,685 target genes. Per-cell enrichment (AUC) of eRegulon genes and regions was calculated using the AUCell function. High confidence eRegulons were selected as those where the absolute correlation between TF expression and gene-based AUC was more than 0.6, the absolute correlation between TF expression and region-based AUC was more than 0.6, and the correlation between gene-based and region-based AUC was more than 0.4.

**Network analysis**

The high-confidence eRegulons were used to construct a network where nodes represent TFs and target genes. The network was constructed using tidygraph (v1.2.2)[82] and visualised using ggraph (v2.1.0)[83]. Centrality metrics were calculated using tidygraph.

**TF footprinting**

In order to carry out TF footprinting for each cluster, the subset-bam tool from 10x (v1.1.0) was used to produce bam files containing reads for cells assigned to each cluster. Individual replicates were merged using samtools [84,85] (v1.16.1) to produce one bam file per cluster. These bam files were used as input to TOBIAS (v0.13.3) [32] for footprinting of motifs from from Vierstra et al. 2020 [23] obtained from https://resources.altius.org/~jvierstra/projects/motif-clustering-v2.0beta/ within peaks identified in any cluster from the Seurat object. Regions falling into the blacklist regions included in the Signac package were excluded. TOBIAS was run in time-series mode on germline clusters only.

TOBIAS PlotAggregate was used to aggregate TF binding scores in 120bp windows across all detected motifs within peaks. For each motif and cluster, the average signal in the central 20bp around the motif and average signal in the flanking regions were calculated. Motifs with a notable footprint have less signal in the central region than in the flanking regions (a positive footprint depth); motifs lacking footprint signal in all clusters were removed from

analysis, leaving 296 motifs of interest. Aggregate binding scores are visualised in Fig 2F and Fig S8C-D.

**DMRT1 ChIP-seq analysis**

DMRT1 ChIP-seq peaks from [33] were downloaded from GEO series GSE64892 and converted to hg38 coordinates using the liftOver function from rtracklayer (v1.56.1)[86], keeping only peaks which lifted over to a single continuous location. The regioneR package (v1.28.0)[87] was used to test for significance of overlap of DMRT1 ChIP-seq peaks and DMRT1 eRegulon target regions compared to all other SCENIC+ eRegulon target regions, using resampleRegions with default parameters and 1000 permutations.

**GWAS data integration**

rsIDs for SNPs associated with infertility in published GWAS studies were obtained from the NHGRI-EBI GWAS catalog [37,38,40] and directly from [36]. The LDLinkR package (v1.2.2)[42] was used to obtain lists of all SNPs within 500kb of the original SNPs and in linkage disequilibrium with them in the respective study population (Han Chinese or European ancestry)[42,43], using genome build "grch38_high_coverage". SNPs with $R^2 > 0.8$ were selected for further analysis.

**TF and target gene changes in cryptozoospermia**

To address the question of whether a change in expression of a TF is associated with changes in expression of its target genes, TFs with a significant ($p < 0.05$) change in expression between normal and cryptozoospermia samples and an absolute log2 fold change in expression of at least 0.5 in at least one cluster were selected. For each TF, its SCENIC+ targets were identified. For each cluster, the mean log2 fold change in expression of targets was calculated. As several clusters show an overall skew in change of gene expression, the mean log2 fold change of all genes in the given cluster was calculated and subtracted from the average change in expression of the target genes. To calculate the significance of the association in direction of change of expression of the TF and target genes, 2x2 contingency tables were calculated per TF and per cluster, for the classes "target/non-target" and "same direction/other direction". Fisher's exact test p-values were calculated and adjusted for multiple testing using the Benjamini-Hochberg method.

**Acknowledgements**

We thank Frank Tüttelmann, Jörg Gromoll, James Turner, Harry Leitch, and Will Scott for advice and comments on the project, and Paul-Georg Majev for advice on data analysis. We also thank Nicole Terwort for assistance with processing of testicular tissues. We thank Heidi Kersebom and Elke Kößer for histological evaluation of testicular tissues and Sabine Forsthoff for excellent support in endocrinological measurements. Finally, we thank the technicians of the andrology laboratory for their support in analysing ejaculate parameters. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) Clinical Research Unit CRU326 'Male Germ Cells: from Genes to Function' (project numbers LA

4064/3-2, NE 2190/3-2 and VA 1456/2 to SL, NN, and JMV respectively). Work in the Vaquerizas laboratory is also supported by the Medical Research Council, UK (award reference MC_UP_160510 to J.M.V.), the Academy of Medical Sciences and the Department of Business, Energy and Industrial Strategy (award reference APR3\1017 to J.M.V.), and the DFG Priority Programme SPP2202 'Spatial Genome Architecture in Development and Disease' (project number VA 1456/1-1 to J.M.V.).

**Author contributions**

EI-S and JB analysed data. JB and SDP produced sn multiome libraries. J-FC and SK carried out surgery, sample and data collection, interpretation and specification of the clinical data set. CF, SL, NN, and JMV obtained funding and supervised the study. EI-S, JB, SDP, NR, CF, SL, NN, and JMV interpreted data. EI-S and JMV wrote the first draft of the manuscript and all authors contributed to editing the manuscript.

**Competing interests**

The authors declare no competing interests.

**Data availability**

Single-nucleus multiome sequencing data will be uploaded to the European Genome-Phenome Archive (EGA). Upon publication, access to the data will be available via application to the Data Access Committee.

**Code availability**

Code will be made available via Github.

# References


1. World Health Organisation. *Infertility prevalence estimates, 1990-2021*. (2023).
2. Agarwal, A. *et al.* Male infertility. *The Lancet* **397**, 319–333 (2021).
3. Tüttelmann, F., Ruckert, C. & Röpke, A. Disorders of spermatogenesis: Perspectives for novel genetic diagnostics after 20 years of unchanged routine. *Med. Genet.* **30**, 12–20 (2018).
4. Mazzilli, R. *et al.* Male factor infertility and assisted reproductive technologies: indications, minimum access criteria and outcomes. *J. Endocrinol. Invest.* **46**, 1079–1085 (2023).
5. Pinborg, A., Wennerholm, U.-B. & Bergh, C. Long-term outcomes for children conceived by assisted reproductive technology. *Fertil. Steril.* (2023) doi:10.1016/j.fertnstert.2023.04.022.
6. Wistuba, J., Stukenborg, J.-B. & Luetjens, C. M. Mammalian Spermatogenesis.
7. O'Donnell, L., Smith, L. B. & Rebourcet, D. Sertoli cells as key drivers of testis function. *Semin. Cell Dev. Biol.* **121**, 2–9 (2022).
8. Heinrich, A. & DeFalco, T. Essential roles of interstitial cells in testicular development and function. *Andrology* **8**, 903–914 (2020).
9. Soumillon, M. *et al.* Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cell Rep.* **3**, 2179–2190 (2013).
10. Xia, B. *et al.* Widespread Transcriptional Scanning in the Testis Modulates Gene Evolution Rates. *Cell* **180**, 248-262.e21 (2020).
11. Di Persio, S. & Neuhaus, N. Human spermatogonial stem cells and their niche in male (in)fertility: novel concepts from single-cell RNA-sequencing. *Hum. Reprod.* deac245 (2022) doi:10.1093/humrep/deac245.
12. Garcia-Alonso, L. *et al.* Single-cell roadmap of human gonadal development. *Nature* 1–8 (2022) doi:10.1038/s41586-022-04918-4.
13. Guo, J. *et al.* The Dynamic Transcriptional Cell Atlas of Testis Development during Human Puberty. *Cell Stem Cell* **26**, 262-276.e4 (2020).
14. Murat, F. *et al.* The molecular evolution of spermatogenesis across mammals. *Nature* 1–9 (2022) doi:10.1038/s41586-022-05547-7.
15. Turner, J. M. A. Meiotic Silencing in Mammals. *Annu. Rev. Genet.* **49**, 395–412 (2015).
16. Wang, T., Gao, H., Li, W. & Liu, C. Essential Role of Histone Replacement and Modifications in Male Fertility. *Front. Genet.* **10**, 962 (2019).
17. Wu, X. *et al.* Single-cell ATAC-Seq reveals cell type-specific transcriptional regulation and unique chromatin accessibility in human spermatogenesis. *Hum. Mol. Genet.* **31**, 321–333 (2022).
18. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
19. Wang, M. *et al.* Single-Cell RNA Sequencing Analysis Reveals Sequential Cell Fate Transition during Human Spermatogenesis. *Cell Stem Cell* **23**, 599-614.e4 (2018).

20. Guo, J. *et al.* The adult human testis transcriptional cell atlas. *Cell Res.* **28**, 1141–1157 (2018).
21. Di Persio, S. *et al.* Single-cell RNA-seq unravels alterations of the human spermatogonial stem cell compartment in patients with impaired spermatogenesis. *Cell Rep. Med.* **2**, 100395 (2021).
22. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
23. Vierstra, J. *et al.* Global reference mapping and dynamics of human transcription factor footprints. *Nature* **583**, (2020).
24. González-Blas, C. B. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. 2022.08.19.504505 Preprint at https://doi.org/10.1101/2022.08.19.504505 (2022).
25. Zhang, T. & Zarkower, D. DMRT proteins and coordination of mammalian spermatogenesis. *Stem Cell Res.* **24**, 195–202 (2017).
26. Zarkower, D. & Murphy, M. W. DMRT1: An Ancient Sexual Regulator Required for Human Gonadogenesis. *Sex. Dev. Genet. Mol. Biol. Evol. Endocrinol. Embryol. Pathol. Sex Determ. Differ.* **16**, 112–125 (2022).
27. Matson, C. K. *et al.* The mammalian doublesex homolog DMRT1 is a transcriptional gatekeeper that controls the mitosis versus meiosis decision in male germ cells. *Dev. Cell* **19**, 612–624 (2010).
28. Peri, A. & Serio, M. The CREM system in human spermatogenesis. *J. Endocrinol. Invest.* **23**, 578–583 (2000).
29. Krausz, C. & Sassone-Corsi, P. Genetic control of spermiogenesis: insights from the CREM gene and implications for human infertility. *Reprod. Biomed. Online* **10**, 64–71 (2005).
30. Pan, J. *et al.* RNF17, a component of the mammalian germ cell nuage, is essential for spermiogenesis. *Dev. Camb. Engl.* **132**, 4029–4039 (2005).
31. Wasik, K. A. *et al.* RNF17 blocks promiscuous activity of PIWI proteins in mouse testes. *Genes Dev.* **29**, 1403–1415 (2015).
32. Bentsen, M. *et al.* ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat. Commun.* **11**, 4267 (2020).
33. Murphy, M. W. *et al.* An ancient protein-DNA interaction underlying metazoan sex determination. *Nat. Struct. Mol. Biol.* **22**, 442–451 (2015).
34. Mesrouze, Y. *et al.* A new perspective on the interaction between the Vg/VGLL1-3 proteins and the TEAD transcription factors. *Sci. Rep.* **10**, 17442 (2020).
35. Figeac, N. *et al.* VGLL3 operates via TEAD1, TEAD3 and TEAD4 to influence myogenesis in skeletal muscle. *J. Cell Sci.* **132**, jcs225946 (2019).
36. Cerván-Martín, M. *et al.* Immune and spermatogenesis-related loci are involved in the development of extreme patterns of male infertility. *Commun. Biol.* **5**, 1–13 (2022).
37. Hu, Z. *et al.* A genome-wide association study in Chinese men identifies three risk loci for non-obstructive azoospermia. *Nat. Genet.* **44**, 183–186 (2012).

38. Zhao, H. *et al.* A Genome-wide Association Study Reveals that Variants within the HLA Region Are Associated with Risk for Nonobstructive Azoospermia. *Am. J. Hum. Genet.* **90**, 900–906 (2012).
39. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).
40. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
41. Gaulton, K. J., Preissl, S. & Ren, B. Interpreting non-coding disease-associated human variants using single-cell epigenomics. *Nat. Rev. Genet.* 1–19 (2023) doi:10.1038/s41576-023-00598-6.
42. Myers, T. A., Chanock, S. J. & Machiela, M. J. LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. *Front. Genet.* **11**, (2020).
43. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
44. Carpentier, M. *et al.* Reduced fertility in male mice deficient in the zinc metallopeptidase NL1. *Mol. Cell. Biol.* **24**, 4428–4437 (2004).
45. Ohsako, T. *et al.* The Drosophila Neprilysin 4 gene is essential for sperm function following sperm transfer to females. *Genes Genet. Syst.* **96**, 177–186 (2021).
46. Lea, R. A. *et al.* KLF17 promotes human naïve pluripotency but is not required for its establishment. *Dev. Camb. Engl.* **148**, dev199378 (2021).
47. Gumireddy, K. *et al.* KLF17 is a negative regulator of epithelial-mesenchymal transition and metastasis in breast cancer. *Nat. Cell Biol.* **11**, 1297–1304 (2009).
48. Robinson, M., Sparanese, S., Witherspoon, L. & Flannigan, R. Human in vitro spermatogenesis as a regenerative therapy — where do we stand? *Nat. Rev. Urol.* 1–19 (2023) doi:10.1038/s41585-023-00723-4.
49. Wyrwoll, M. J. *et al.* Bi-allelic Mutations in M1AP Are a Frequent Cause of Meiotic Arrest and Severely Impaired Spermatogenesis Leading to Male Infertility. *Am. J. Hum. Genet.* **107**, 342–351 (2020).
50. Wyrwoll, M. J. *et al.* Genetic Architecture of Azoospermia—Time to Advance the Standard of Care. *Eur. Urol.* **83**, 452–462 (2023).
51. Li, Y. *et al.* M1AP interacts with the mammalian ZZS complex and promotes male meiotic recombination. *EMBO Rep.* **n/a**, e55778 (2022).
52. Groza, T. *et al.* The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease. *Nucleic Acids Res.* **51**, D1038–D1045 (2023).
53. Ichiyanagi, T. *et al.* HSP90α plays an important role in piRNA biogenesis and retrotransposon repression in mouse. *Nucleic Acids Res.* **42**, 11903–11911 (2014).
54. Kherraf, Z.-E. *et al.* SPINK2 deficiency causes infertility by inducing sperm defects in heterozygotes and azoospermia in homozygotes. *EMBO Mol. Med.* **9**, 1132–1149 (2017).

55. Cong, J. *et al.* Homozygous mutations in CCDC34 cause male infertility with oligoasthenoteratozoospermia in humans and mice. *J. Med. Genet.* **59**, 710–718 (2022).
56. Ruggiu, M. *et al.* The mouse Dazla gene encodes a cytoplasmic protein essential for gametogenesis. *Nature* **389**, 73–77 (1997).
57. Gill, M. E., Hu, Y.-C., Lin, Y. & Page, D. C. Licensing of gametogenesis, dependent on RNA binding protein DAZL, as a gateway to sexual differentiation of fetal germ cells. *Proc. Natl. Acad. Sci.* **108**, 7443–7448 (2011).
58. Li, H. *et al.* DAZL is a master translational regulator of murine spermatogenesis. *Natl. Sci. Rev.* **6**, 455–468 (2019).
59. Shin, Y.-H. *et al.* Hormad1 mutation disrupts synaptonemal complex formation, recombination, and chromosome segregation in mammalian meiosis. *PLoS Genet.* **6**, e1001190 (2010).
60. Okutman, O., Boivin, M., Muller, J., Charlet-Berguerand, N. & Viville, S. A biallelic loss of function variant in HORMAD1 within a large consanguineous Turkish family is associated with spermatogenic arrest. *Hum. Reprod. Oxf. Engl.* **38**, 306–314 (2023).
61. Fang, X. *et al.* Hypomorphic and hypermorphic mouse models of Fsip2 indicate its dosage-dependent roles in sperm tail and acrosome formation. *Development* **148**, dev199216 (2021).
62. Martinez, G. *et al.* Whole-exome sequencing identifies mutations in FSIP2 as a recurrent cause of multiple morphological abnormalities of the sperm flagella. *Hum. Reprod.* **33**, 1973–1984 (2018).
63. Zheng, R. *et al. FSIP2* plays a role in the acrosome development during spermiogenesis. *J. Med. Genet.* **60**, 254–264 (2023).
64. Szedlak, A., Smith, N., Liu, L., Paternostro, G. & Piermarocchi, C. Evolutionary and Topological Properties of Genes and Community Structures in Human Gene Regulatory Networks. *PLoS Comput. Biol.* **12**, (2016).
65. Alvarez-Ponce, D., Feyertag, F. & Chakraborty, S. Position Matters: Network Centrality Considerably Impacts Rates of Protein Evolution in the Human Protein–Protein Interaction Network. *Genome Biol. Evol.* **9**, 1742–1756 (2017).
66. McDonald, J. M. C. & Reed, R. D. Patterns of selection across gene regulatory networks. *Semin. Cell Dev. Biol.* **145**, 60–67 (2023).
67. Kitano, H. Biological robustness. *Nat. Rev. Genet.* **5**, 826–837 (2004).
68. Cerván-Martín, M., Castilla, J. A., Palomino-Morales, R. J. & Carmona, F. D. Genetic Landscape of Nonobstructive Azoospermia and New Perspectives for the Clinic. *J. Clin. Med.* **9**, 300 (2020).
69. Kossack, N. *et al.* A combined approach facilitates the reliable detection of human spermatogonia in vitro. *Hum. Reprod. Oxf. Engl.* **28**, 3012–3025 (2013).
70. Laurentino, S. *et al.* High-resolution analysis of germ cells from men with sex chromosomal aneuploidies reveals normal transcriptome but impaired imprinting. *Clin. Epigenetics* **11**, 127 (2019).

71. Neuhaus, N. *et al.* Single-cell gene expression analysis reveals diversity among human spermatogonia. *Mol. Hum. Reprod.* **23**, 79–90 (2017).
72. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
73. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience* **9**, (2020).
74. Germain, P.-L., Lun, A., Garcia Meixide, C., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Research* **10**, 979 (2021).
75. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* btw777 (2017) doi:10.1093/bioinformatics/btw777.
76. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
77. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5**, 2122 (2016).
78. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
79. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS J. Integr. Biol.* **16**, 284–287 (2012).
80. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
81. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
82. Pedersen, T. L. *tidygraph: A Tidy API for Graph Manipulation*. (2023).
83. Pedersen, T. L. *ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. (2022).
84. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
85. Bonfield, J. K. *et al.* HTSlib: C library for reading/writing high-throughput sequencing data. *GigaScience* **10**, giab007 (2021).
86. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).
87. Gel, B. *et al.* regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32**, 289–291 (2016).

# Figure Legends

**Figure 1: Single-nucleus gene expression and chromatin accessibility characterisation of human spermatogenesis**

A. Schematic overview of the stages of human spermatogenesis. Undifferentiated spermatogonia commit to differentiation, then become spermatocytes which go through meiosis to produce haploid spermatids. The seminiferous tubules contain germline and Sertoli cells, are surrounded by peritubular myoid cells (PTMs), and the interstitial space between tubules contains Leydig cells, pericytes, endothelial cells, and immune cells. Seminiferous tubules of different stages include different subsets of germline developmental stages; the schematic tubule shown here includes segments of different stages in order to show all germline cell types.

B. UMAP representation of single-nucleus multiomics data from human testis. Inset shows trajectory inference performed on germline clusters using Slingshot.

C. Heatmap of gene expression across spermatogenesis. Clusters are ordered by their position in the trajectory shown in B, followed by somatic cells. Genes are ordered by the cluster in which the gene has max expression. Selected marker genes for different cell types are labelled on the right-hand side.

D. Examples of dynamic chromatin accessibility at selected marker genes. Each sub-panel shows cluster-specific pseudobulk chromatin accessibility tracks on the left, with gene annotations below. Regions are centred on genes of interest, which are labelled. Expression of the marker gene of interest in the cluster is shown in the dot plots on the right, where colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression. The *TNP1* locus contains a highly accessible putative enhancer (left) in addition to gene-associated accessibility.

E. Heatmap of motif activity across spermatogenesis. Clusters are ordered by their position in the trajectory shown in B, followed by somatic cells. Motifs are ordered by the cluster in which the motif has highest activity. The transcription factor families for selected cluster-specific motifs are shown on the right-hand side.

**Figure 2: SCENIC+ identifies stage-specific eRegulons and indicates differences in regulation between spermatids and other stages**

A. Schematic depiction of eRegulon identification by SCENIC+. SCENIC+ identifies putative enhancers containing transcription factor (TF) motifs. TFs, putative enhancer regions (R1-R3), and target genes (G1-G3) are grouped into eRegulons on the basis of correlation between the TF expression, target gene expression, and region accessibility. An eRegulon consists of a TF, its target regions, and its target genes.

B. 62 high quality eRegulons which show dynamic activity across cell types were identified. Top, activator eRegulons. Bottom, repressor eRegulons. Left-to-right, panels represent eRegulon target gene expression enrichment per cluster (blue, scaled average AUC), eRegulon target region accessibility enrichment per cluster (green, scaled average AUC), and TF expression per cluster (magenta, Z-score). There are 4 distinct groups of eRegulons active in spermatogonia (1), differentiating spermatogonia and spermatocytes (2), spermatids (3) and somatic cell types (4).

C. The gene regulatory network made up of the 62 identified high quality eRegulons. Nodes represent genes, coloured by the cluster in which the gene has max expression. Colour assignments are as in Fig 1. TFs are labelled and shown as large diamonds and all other genes as smaller circles. Label background colour is white or grey for readability purposes only. Edges represent a regulatory relationship between a TF and a target gene inferred by SCENIC+.
D. Quantification of network properties for activating TFs. TFs driving eRegulons in Group 3, active in spermatids, have significantly lower betweenness centrality and in degree compared to TFs active in Group 1 or Group 4. Significance asterisks are shown for pairwise comparisons of Group 3 with all other groups; **: $p <= 0.01$, *: $p <= 0.05$, ns: $p > 0.05$.
E. SCENIC+ identifies two predicted enhancers for *RNF17* (red bars within grey highlights). One of the enhancers is accessible in spermatogonia, the other primarily in zygotene and pachytene/diplotene, as shown by cluster-specific pseudobulk chromatin accessibility tracks.
F. Selected motifs targeted by LIN28B (top), and CLOCK (bottom). The motif AC0069 is part of a cluster of similar motifs bound by MAX and HES family TFs in addition to CLOCK. The TOBIAS footprint profiles of these motifs indicate that they are bound most strongly in pachytene/diplotene.

**Figure 3: SCENIC+ results identify candidate target genes and enhancers for GWAS SNPs.**

A. *MMEL1* locus. Left, cluster-specific pseudobulk chromatin accessibility tracks. Right, *MMEL1* expression per cluster. Bottom, gene annotations, SCENIC+ predicted distal enhancers for *MMEL1*, SNPs, and snATAC-seq peaks. The original SNP (dark blue) is rs2477686 identified as associated with male infertility by Hu et al. 2012; proxy SNPs in LD ($r2 > 0.8$) with this SNP are shown in light blue.
B. Top, close up of part of the region in A. Two proxy SNPs of rs2477686 overlap an snATAC-seq peak identified by SCENIC+ as a putative enhancer for *MMEL1*, regulated by KLF17.
C. KLF17 eRegulon activity across single nuclei, as measured by target gene expression enrichment (top) and target region chromatin accessibility enrichment (bottom).
D. *KLF17* expression across clusters; colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression.

**Figure 4: Single-cell multiomics of spermatogenesis in infertile men**

A. UMAP representation of single-nucleus multiomics data from normal spermatogenesis (left, combined samples), and two infertile individuals with cryptozoospermia.
B. Quantification of the contribution of each biological replicate to the clusters shown in A. The samples from infertile individuals have a noticeable depletion in the number of spermatids, and increase in Leydig cells and one subcluster of PTMs.
C. Numbers of differentially expressed genes (left) and differentially accessible peaks (right) in each cluster, comparing normal spermatogenesis and Crypto 2.
D. Differentially expressed genes between normal spermatogenesis and Crypto 2 shown in the context of the high-confidence gene regulatory network (Fig 2C). TFs are shown as large diamonds and all other genes as smaller circles. Genes that are significantly differentially expressed in any cluster are shown in red and all other genes in grey. Edges represent a regulatory relationship between a TF and a target gene inferred by SCENIC+. Differentially

expressed TFs are labelled and coloured according to the cluster in which they have highest expression in the normal spermatogenesis samples.
E. The target genes of significantly differentially expressed SCENIC+ TFs show changes in gene expression consistent with SCENIC+ predictions. Example of the distribution of log2FC in expression between Crypto 2 and normal samples, for all genes (grey) and DMRTB1 target genes (red). The mean of the distribution is shown with a corresponding vertical dashed line. The log2FC of DMRTB1 is shown with a black vertical dashed line. The difference in mean log2FC between target genes and all genes is plotted on the y axis in panel F.
F. The target genes of significantly differentially expressed SCENIC+ TFs show changes in gene expression consistent with SCENIC+ predictions. Activator TFs significantly differentially expressed in any cluster between normal and Crypto 2 ($p < 0.05$, absolute log2FC > 1) are shown. Each point represents a TF and its targets in a given cluster; the logFC of the TF is represented on the x axis and the difference in mean log2FC between target genes and all genes is plotted on the y axis, as explained in panel E. Clusters are coloured according to the key in A. Point size reflects the Fisher test p-value of the association between direction of change in expression for the TF and direction of change in expression of its targets.
G. Examples of genes that are downregulated in Crypto 2 compared to normal spermatogenesis samples. Average normalised expression per cluster is shown for normal (solid line) and Crypto 2 (dotted line). Colours correspond to germline clusters according to the key in panel A.

# Supplementary Figure Legends

**Figure S1: Identification of clusters**
A. WNN UMAP showing all clusters, including 4 “Unclear” clusters that could not be identified based on marker gene expression and were excluded from further analysis (highlighted with dashed circles).
B. Marker gene expression across clusters. Colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression.
C. Gene Ontology (Biological Process) terms that are significantly enriched in marker genes for each cluster shown in Fig 1. The top most significant term per cluster was selected for inclusion in the plot. Size represents -log10 of the BH-adjusted p-value for enrichment of the term in the given gene set, and colour represents the number of DE genes associated with that term. Absent dots indicate no significant enrichment of that term in the given gene set.

**Figure S2: RNA/ATAC-seq integration**
A. UMAPs based on RNA-seq, ATAC-seq, and the weighted nearest neighbour (WNN) graph produced by combining the two modalities using Seurat.
B. RNA/ATAC-seq weight per cluster. Most clusters have a 50:50 weighting of RNA and ATAC-seq; the ‘Unclear’ clusters are skewed towards 100% weighting of one modality.

**Figure S3: Characteristics of ATAC-seq signal across clusters**

A. The fraction of all ATAC-seq reads in peaks coming from chrX is consistent throughout spermatogonia and in somatic cells, but drops in zygotene and pachytene/diplotene, and appears bimodal in meiotic divisions and spermatids, indicative of haploid cells containing either the X or Y chromosome.
B. The fraction of reads per cell which fall inside peaks drops sharply at the meiotic divisions stage and in spermatids. Cells with less than 20% of reads in peaks were filtered out during pre-processing as these are likely to be low quality cells.

**Figure S4: Comparison of snRNA-seq and scRNA-seq.**
A. UMAPs of clusters obtained from snRNA-seq (this study) and published scRNA-seq from Di Persio & Tekath et al. 2021 and Guo et al. 2018[20,21].
B. Proportion of different cell types captured across the three datasets indicating an enrichment of spermatogonia and a depletion of spermatids in snRNA-seq compared to scRNA-seq datasets.

**Figure S5: Further examples of dynamic chromatin accessibility at cluster-specific marker genes.**
Each sub-panel shows cluster-specific pseudobulk chromatin accessibility tracks on the left. Expression of the marker gene of interest in the cluster shown in the dot plots on the right, where colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression. Marker genes shown are *CD14* for macrophages, *NKG7* for T cells, *PECAM1* for endothelial cells, *DDX4* for germline cells, *VIM* for somatic cells, and *SOX9* for Sertoli cells.

**Figure S6: Heatmap of eRegulon activity across clusters.**
62 high quality eRegulons which show dynamic activity across cell types were identified. TF name and direction of regulation are labelled on the right. Top, activator eRegulons. Bottom, repressor eRegulons. Left-to-right, panels represent eRegulon target gene expression enrichment per cluster (blue, scaled average AUC), eRegulon target region accessibility enrichment per cluster (green, scaled average AUC), and TF expression per cluster (magenta, Z-score). There are 4 distinct clusters of eRegulons active in spermatogonia (1), differentiating spermatogonia and spermatocytes (2), spermatids (3) and somatic cell types (4).

**Figure S7: Further examples of SCENIC+ predicted enhancers and target genes.**
A. SCENIC+ identifies an enhancer (red bar within grey highlight) upstream of *RMND5A* which is regulated by DMRT1, DMRTB1, and FOXO1. Left, cluster-specific pseudobulk chromatin accessibility tracks; right, *RMND5A* expression shown as a dotplot where colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression. Bottom, gene annotations, predicted enhancers (red), and scATAC-seq peaks (grey). A DMRT1 ChIP-seq peak overlapping the predicted enhancer is shown with a blue bar.
B. SCENIC+ identifies two enhancers for *VGLL3* (red bars within grey highlights). The lefthand enhancer is predicted to be regulated by EBF1 and NR2F2 and the righthand enhancer by ATF5 and CREM. Left, cluster-specific pseudobulk chromatin accessibility tracks; right, *VGLL3* expression shown as a dotplot where colour indicates scaled average

expression and size indicates the % of cells in the cluster with detected expression. Bottom, gene annotations, predicted enhancers (red), and scATAC-seq peaks (grey).
C. TF footprinting for motifs bound by DMRT1/DMRTB1 (top) and FOXO1 (bottom), showing highest activity in differentiating spermatogonia.
D. TF footprinting for motifs bound by NR2F2 (top), showing little activity in germline cells, and by ATF5/CREM (bottom), showing highest activity in late spermatids.

**Figure S8: SCENIC+ results identify *TNFRSF14* as a candidate target gene for a GWAS SNPs associated with infertility.**
A. *TNFRSF14* locus. Left, cluster-specific pseudobulk chromatin accessibility tracks. Right, *TNFRSF14* expression per cluster as a dotplot; colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression. Bottom, gene annotations, SCENIC+ predicted distal enhancers for *TNFRSF14*, SNPs, and snATAC-seq peaks. The original SNP (dark blue) is rs2477686 identified as associated with male infertility by Hu et al. 2012; proxy SNPs in LD ($r2 > 0.8$) with this SNP are shown in light blue.
B. Top, close up of part of the region in A. A proxy SNP of rs2477686 overlaps an snATAC-seq peak identified by SCENIC+ as a putative enhancer for *TNFRSF14*, regulated by ATF3, EBF1, FOXC1, and NFIC.
C. Expression of *ATF3*, *EBF1*, *FOXC1*, and *NFIC* as a dotplot; colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression.
D, E. eRegulon activity for ATF3, EBF1, FOXC1, and NFIC across clusters, as measured by target gene expression enrichment (D) and target region chromatin accessibility enrichment (E).

**Fig S9: Identification of clusters in the combined data set (normal + cryptozoospermia)**
Marker gene expression across clusters. Colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression.

**Fig S10: Differential gene expression and chromatin accessibility in Crypto 1**
A. Numbers of differentially expressed genes (left) and differentially accessible peaks (right) in each cluster, comparing normal spermatogenesis and Crypto 1.
B. Differentially expressed genes between normal spermatogenesis and Crypto 1 shown in the context of the high-confidence gene regulatory network (Fig 2C). TFs are shown as large diamonds and all other genes as smaller circles. Genes that are significantly differentially expressed in any cluster are shown in red and all other genes in grey. Edges represent a regulatory relationship between a TF and a target gene inferred by SCENIC+. Differentially expressed TFs are labelled and coloured according to the cluster in which they have highest expression in the normal spermatogenesis samples.
C. The target genes of significantly differentially expressed SCENIC+ TFs show changes in gene expression consistent with SCENIC+ predictions. Activator TFs significantly differentially expressed in any cluster between normal and Crypto 1 ($p < 0.05$, absolute $log2FC > 1$) are shown. Each point represents a TF and its targets in a given cluster; the logFC of the TF is represented on the x axis and the difference in mean log2FC between target genes and all genes is plotted on the y axis, as explained in Fig 4E. Clusters are coloured according to the key in A. Point size reflects the Fisher test p-value of the

association between direction of change in expression for the TF and direction of change in expression of its targets.
D. As in C, for repressive TFs significantly differentially expressed in any cluster between normal and Crypto 1 ($p < 0.05$, absolute log2FC > 1).
E. As in C, for repressive TFs significantly differentially expressed in any cluster between normal and Crypto 2 ($p < 0.05$, absolute log2FC > 1).

**Fig S11: Enriched GO terms for differentially expressed genes in Crypto 2**
The most significant GO term for each cluster is shown, considering up- and downregulated genes in each cluster separately. Size represents -log10 of the BH-adjusted p-value for enrichment of the term in the given gene set, and colour represents the number of DE genes associated with that term. Absent dots indicate no significant enrichment of that term in the given gene set. BP: biological process; CC: cellular component; MF: molecular function.

# Supplementary Tables

**Table S1 – Clinical parameters for donors with normal spermatogenesis**

FSH, follicle-stimulating hormone; LH, luteinising hormone; SHBG, sex hormone-binding globulin; DHT, dihydrotestosterone; TESE, testicular sperm extraction; CFTR, cystic fibrosis transmembrane conductance regulator gene; AZF, azoospermia factor.

| | **Sample 1** | **Sample 2** | **Sample 3** |
|---|---|---|---|
| **Age (Years)** | 32 | 33 | 41 |
| **Ejaculate Volume (ml)** | 4.6 | 0.7 | 4.7 |
| **Total Sperm Count (Millions)** | 0 | 0 | 0 |
| **Sperm Concentration (Million/ml)** | 0 | 0 | 0 |
| **FSH (normal range 1-7 U/l)** | 1.2 | 2.2 | 3.7 |
| **LH (normal range 2-10 U/l)** | 4.2 | 2 | 2.3 |
| **Total Testosterone (normal range >12 nmol/l)** | 34.5 | 24.9 | 18 |
| **SHBG (normal range 11-71 nmol/l)** | 51 | 55 | Not available |
| **Free Testosterone (normal range >250 pmol/l)** | 637 | 400 | Not available |
| **Prolactin (normal range <500 mU/l)** | 107 | 142 | Not available |
| **Estradiol (normal range <250 pmol/l)** | 88 | 82 | Not available |
| **DHT (normal range 0.5-2.0 nmol/l)** | 1.67 | 0.55 | Not available |
| **Bitesticular volume (ml)** | 53 | 40 | 38 |
| **Bergmann-Kliesch Score** | 7 | 8 | 8 |
| **% tubules with elongated spermatids** | 66 | 79 | 79 |
| **% tubules with round spermatids** | 1 | 0 | 0 |
| **% tubules with spermatocytes** | 26 | 12 | 13 |
| **% tubules with spermatogonia** | 6 | 2 | 1 |
| **% tubules sertoli cell only** | 1 | 0 | 2 |
| **% tubular shadows** | 0 | 7 | 5 |
| **Retrieval of sperm from the TESE** | Yes | Yes | Yes |
| **Karyotype** | 46,XY | 46,XY | Not available |
| **CFTR genotype** | Not available | Heterozygous for F508del [c.1521_1523delCTT p.Phe508del], pathogen Poly(T)-Allele in Intron 8: T7/T9 | Not available |

| | | | |
|---|---|---|---|
| **AZF deletion** | No deletion | No deletion | Not available |

**Table S2 – High confidence TF-target gene pairs identified by SCENIC+**

Supplied as supplementary file TableS2_selected_TFs_and_targets.xlsx

**Table S3 – SNPs significantly associated with male infertility in published studies and their relationship with scATAC-seq peaks and candidate enhancers**

Note that rs1136759 has no identified proxy SNPs as it is not present in the database used to identify these.

| rsID | Study | Nearest gene | Coding? | Overlaps ATAC-seq peak? | # proxy SNPs | # proxy SNPs overlapping a peak | # proxy SNPs overlapping a putative enhancer |
|---|---|---|---|---|---|---|---|
| rs1136759 | Cerván-Martín et al. 2022 | *HLA-DRB1* | Yes | Yes | N/A | N/A | N/A |
| rs115054029 | Cerván-Martín et al. 2022 | *VRK1* | No | No | 19 | 0 | N/A |
| rs10842262 | Hu et al. 2012 | *SOX5* | No | No | 13 | 0 | N/A |
| rs12097821 | Hu et al. 2012 | *MTATP6P14* | No | No | 147 | 2 | 0 |
| rs2477686 | Hu et al. 2012 | *PLCH2* | No | No | 14 | 4 | 3 |
| rs6080550 | Hu et al. 2012 | *CKAP2LP1* | No | No | 3 | 0 | N/A |
| rs7192 | Zhao et al. 2012 | *HLA-DRA* | No | No | 18 | 0 | N/A |

**Table S4 – Clinical parameters for donors with cryptozoospermia**

FSH, follicle-stimulating hormone; LH, luteinising hormone; SHBG, sex hormone-binding globulin; DHT, dihydrotestosterone; TESE, testicular sperm extraction; CFTR, cystic fibrosis transmembrane conductance regulator gene; AZF, azoospermia factor.

| | **Crypto 1** | **Crypto 2** |
|---|---|---|
| **Age (Years)** | 28 | 42 |
| **Ejaculate Volume (ml)** | 3 | 1.4 |
| **Total Sperm Count (Millions)** | 0 | <0,1 |
| **Sperm Concentration (Million/ml)** | 0 | <0,1 |
| **FSH (normal range 1-7 U/l)** | 4.1 | 8.9 |
| **LH (normal range 2-10 U/l)** | 3.5 | 5.6 |
| **Total Testosterone (normal range >12 nmol/l)** | 17.9 | 17.6 |
| **SHBG (normal range 11-71 nmol/l)** | 25 | 49 |
| **Free Testosterone (normal range >250 pmol/l)** | 444 | 288 |
| **Prolactin (normal range <500 mU/l)** | 226 | 614 |
| **Estradiol (normal range <250 pmol/l)** | 80 | 60 |
| **DHT (normal range 0.5-2.0 nmol/l)** | 0.78 | 0.61 |
| **Bitesticular volume (ml)** | 27 | 30 |
| **Bergmann-Kliesch Score** | 1 | 0.9 |
| **% tubules with elongated spermatids** | 10 | 9 |
| **% tubules with round spermatids** | 10 | 8 |
| **% tubules with spermatocytes** | 76 | 70 |
| **% tubules with spermatogonia** | 3 | 5 |
| **% tubules Sertoli cell only** | 0 | 0 |
| **% tubular shadows** | 2 | 9 |
| **Retrieval of sperm from the TESE** | Yes | No |
| **Karyotype** | 46,XY | 46,XY |
| **CFTR genotype** | Not available | Not available |
| **AZF deletion** | No deletion | No deletion |

# Figures

1138 **Figure 1**

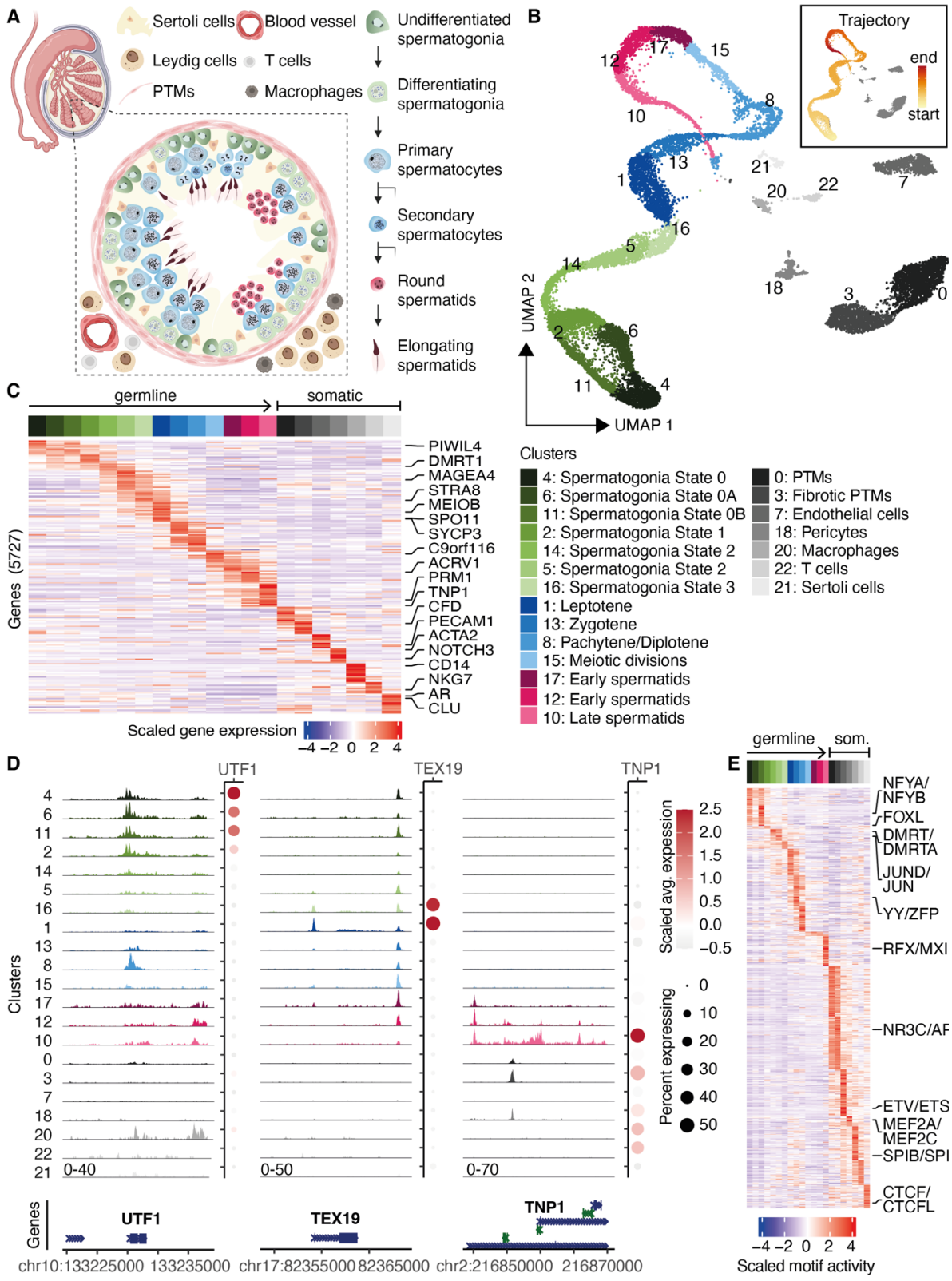



**Figure 1: Single-nucleus gene expression and chromatin accessibility characterisation of human spermatogenesis**

A. Schematic overview of the stages of human spermatogenesis. Undifferentiated spermatogonia commit to differentiation, then become spermatocytes which go through

meiosis to produce haploid spermatids. The seminiferous tubules contain germline and Sertoli cells, are surrounded by peritubular myoid cells (PTMs), and the interstitial space between tubules contains Leydig cells, pericytes, endothelial cells, and immune cells. Seminiferous tubules of different stages include different subsets of germline developmental stages; the schematic tubule shown here includes segments of different stages in order to show all germline cell types.
B. UMAP representation of single-nucleus multiomics data from human testis. Inset shows trajectory inference performed on germline clusters using Slingshot.
C. Heatmap of gene expression across spermatogenesis. Clusters are ordered by their position in the trajectory shown in B, followed by somatic cells. Genes are ordered by the cluster in which the gene has max expression. Selected marker genes for different cell types are labelled on the right-hand side.
D. Examples of dynamic chromatin accessibility at selected marker genes. Each sub-panel shows cluster-specific pseudobulk chromatin accessibility tracks on the left, with gene annotations below. Regions are centred on genes of interest, which are labelled. Expression of the marker gene of interest in the cluster shown in the dot plots on the right, where colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression. The *TNP1* locus contains a highly accessible putative enhancer (left) in addition to gene-associated accessibility.
E. Heatmap of motif activity across spermatogenesis. Clusters are ordered by their position in the trajectory shown in B, followed by somatic cells. Motifs are ordered by the cluster in which the motif has highest activity. The transcription factor families for selected cluster-specific motifs are shown on the right-hand side.

1139

1140 **Figure 2**

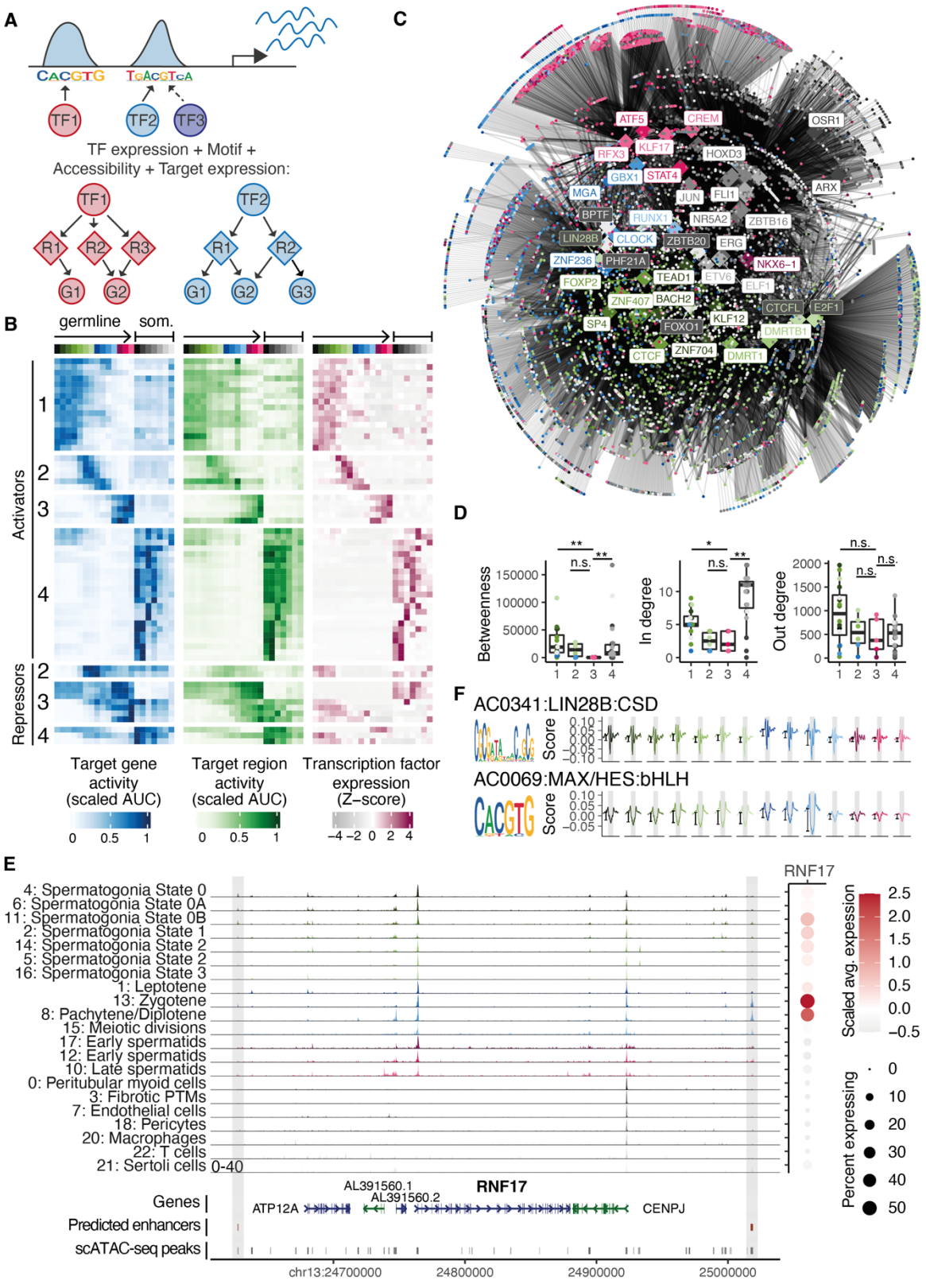



**Figure 2: SCENIC+ identifies stage-specific eRegulons and indicates differences in regulation between spermatids and other stages**

A. Schematic depiction of eRegulon identification by SCENIC+. SCENIC+ identifies putative enhancers containing transcription factor (TF) motifs. TFs, putative enhancer regions (R1-R3), and target genes (G1-G3) are grouped into eRegulons on the basis of correlation between the TF expression, target gene expression, and region accessibility. An eRegulon consists of a TF, its target regions, and its target genes.
B. 62 high quality eRegulons which show dynamic activity across cell types were identified. Top, activator eRegulons. Bottom, repressor eRegulons. Left-to-right, panels represent eRegulon target gene expression enrichment per cluster (blue, scaled average AUC), eRegulon target region accessibility enrichment per cluster (green, scaled average AUC), and TF expression per cluster (magenta, Z-score). There are 4 distinct groups of eRegulons active in spermatogonia (1), differentiating spermatogonia and spermatocytes (2), spermatids (3) and somatic cell types (4).
C. The gene regulatory network made up of the 62 identified high quality eRegulons. Nodes represent genes, coloured by the cluster in which the gene has max expression. Colour assignments are as in Fig 1. TFs are labelled and shown as large diamonds and all other genes as smaller circles. Label background colour is white or grey for readability purposes only. Edges represent a regulatory relationship between a TF and a target gene inferred by SCENIC+.
D. Quantification of network properties for activating TFs. TFs driving eRegulons in Group 3, active in spermatids, have significantly lower betweenness centrality and in degree compared to TFs in Group 1 or Group 4. Significance asterisks are shown for pairwise comparisons of Group 3 with all other groups; **: $p <= 0.01$, *: $p <= 0.05$, ns: $p > 0.05$.
E. SCENIC+ identifies two predicted enhancers for *RNF17* (red bars within grey highlights). One of the enhancers is accessible in spermatogonia, the other primarily in zygotene and pachytene/diplotene, as shown by cluster-specific pseudobulk chromatin accessibility tracks.
F. Selected motifs targeted by LIN28B (top), and CLOCK (bottom). The motif AC0069 is part of a cluster of similar motifs bound by MAX and HES family TFs in addition to CLOCK. The TOBIAS footprint profiles of these motifs indicate that they are bound most strongly in pachytene/diplotene.

1141

1142 **Figure 3**

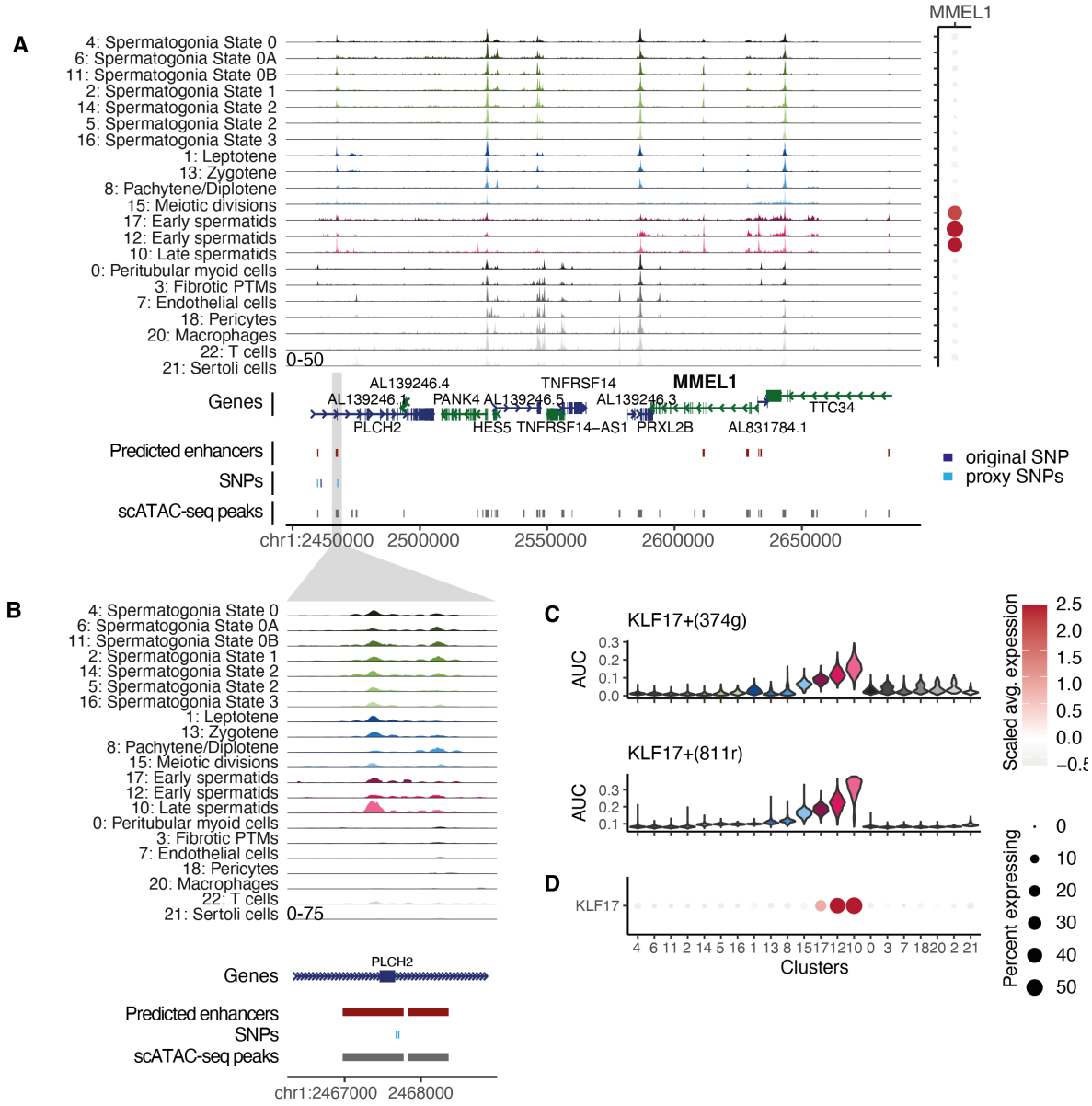



**Figure 3: SCENIC+ results identify candidate target genes and enhancers for GWAS SNPs.**

A. *MMEL1* locus. Left, cluster-specific pseudobulk chromatin accessibility tracks. Right, *MMEL1* expression per cluster. Bottom, gene annotations, SCENIC+ predicted distal enhancers for *MMEL1*, SNPs, and snATAC-seq peaks. The original SNP (dark blue) is rs2477686 identified as associated with male infertility by Hu et al. 2012; proxy SNPs in LD ($r2 > 0.8$) with this SNP are shown in light blue.

B. Top, close up of part of the region in A. Two proxy SNPs of rs2477686 overlap an snATAC-seq peak identified by SCENIC+ as a putative enhancer for *MMEL1*, regulated by KLF17.

C. KLF17 eRegulon activity across single nuclei, as measured by target gene expression enrichment (top) and target region chromatin accessibility enrichment (bottom).

D. *KLF17* expression across clusters; colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression.

1143 **Figure 4**

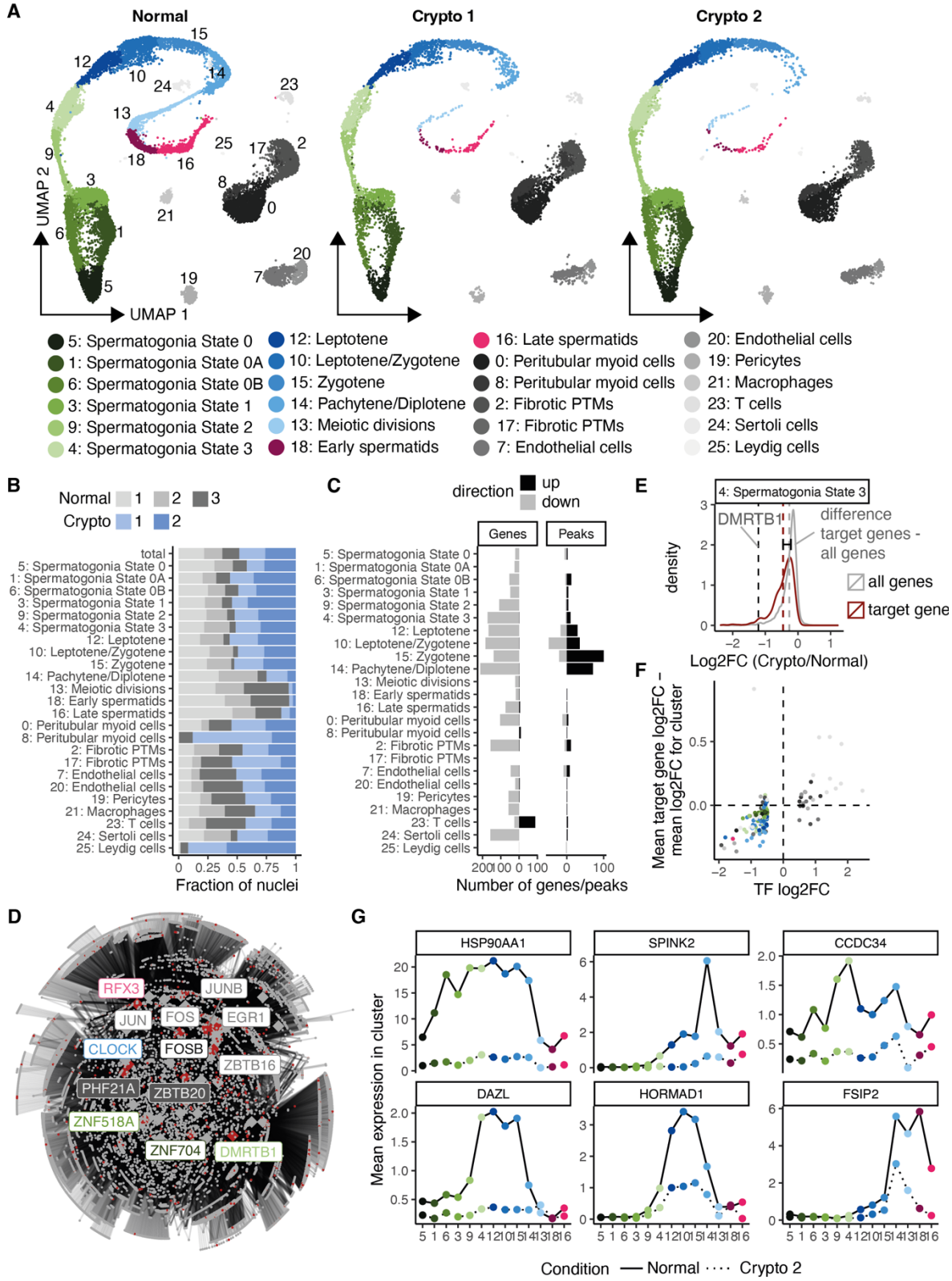



**Figure 4: Single-cell multiomics of spermatogenesis in infertile men**

A. UMAP representation of single-nucleus multiomics data from normal spermatogenesis (left, combined samples), and two infertile individuals with cryptozoospermia.

B. Quantification of the contribution of each biological replicate to the clusters shown in A. The samples from infertile individuals have a noticeable depletion in the number of spermatids, and increase in Leydig cells and one subcluster of PTMs.
C. Numbers of differentially expressed genes (left) and differentially accessible peaks (right) in each cluster, comparing normal spermatogenesis and Crypto 2.
D. Differentially expressed genes between normal spermatogenesis and Crypto 2 shown in the context of the high-confidence gene regulatory network (Fig 2C). TFs are shown as large diamonds and all other genes as smaller circles. Genes that are significantly differentially expressed in any cluster are shown in red and all other genes in grey. Edges represent a regulatory relationship between a TF and a target gene inferred by SCENIC+. Differentially expressed TFs are labelled and coloured according to the cluster in which they have highest expression in the normal spermatogenesis samples.
E. The target genes of significantly differentially expressed SCENIC+ TFs show changes in gene expression consistent with SCENIC+ predictions. Example of the distribution of log2FC in expression between Crypto 2 and normal samples, for all genes (grey) and DMRTB1 target genes (red). The mean of the distribution is shown with a corresponding vertical dashed line. The log2FC of DMRTB1 is shown with a black vertical dashed line. The difference in mean log2FC between target genes and all genes is plotted on the y axis in panel F.
F. The target genes of significantly differentially expressed SCENIC+ TFs show changes in gene expression consistent with SCENIC+ predictions. Activator TFs significantly differentially expressed in any cluster between normal and Crypto 2 ($p < 0.05$, absolute log2FC > 1) are shown. Each point represents a TF and its targets in a given cluster; the logFC of the TF is represented on the x axis and the difference in mean log2FC between target genes and all genes is plotted on the y axis, as explained in panel E. Clusters are coloured according to the key in A. Point size reflects the Fisher test p-value of the association between direction of change in expression for the TF and direction of change in expression of its targets.
G. Examples of genes that are downregulated in Crypto 2 compared to normal spermatogenesis samples. Average normalised expression per cluster is shown for normal (solid line) and Crypto 2 (dotted line). Colours correspond to germline clusters according to the key in panel A.

1144

**Figure S1**

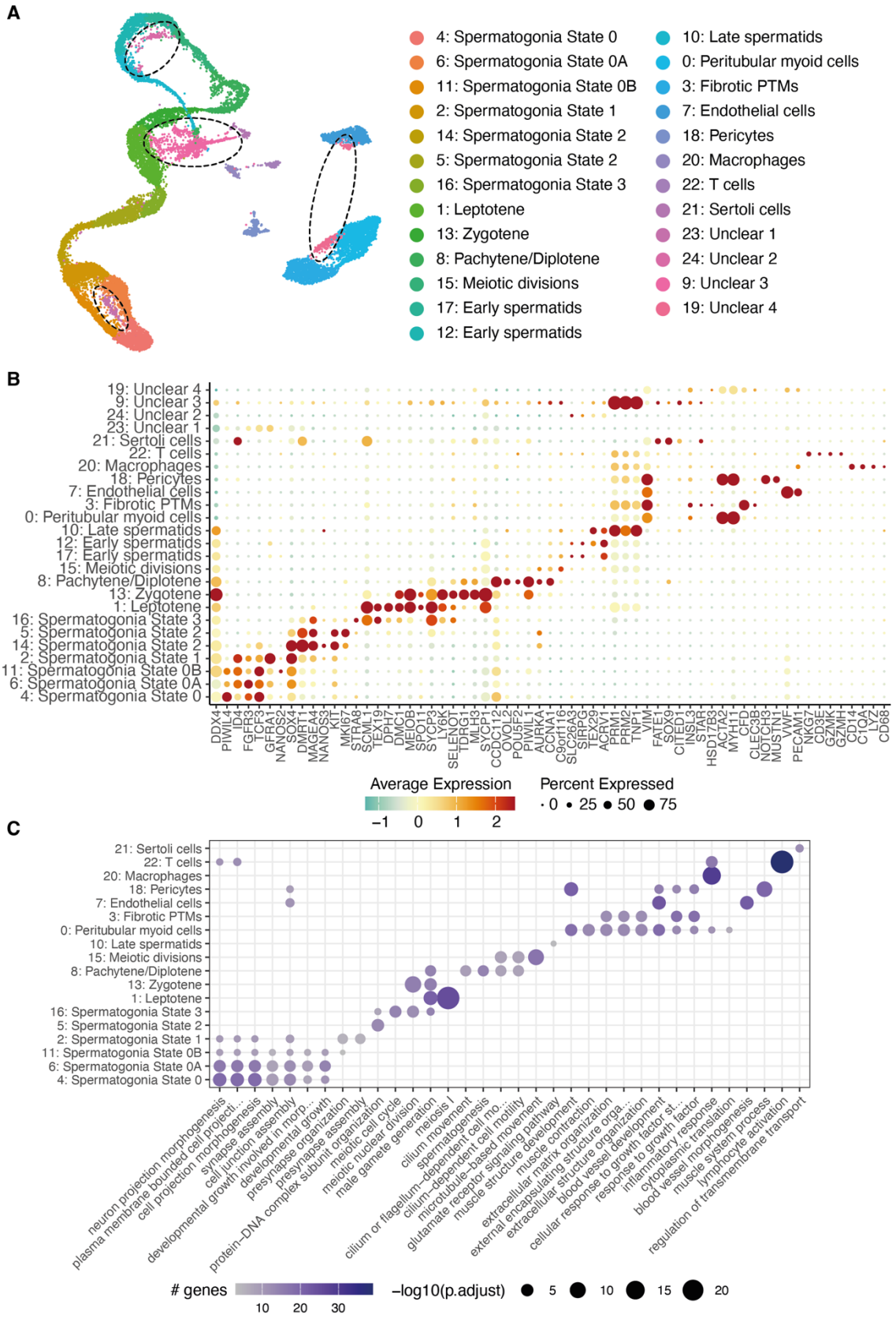



**Figure S1: Identification of clusters**

A. WNN UMAP showing all clusters, including 4 “Unclear” clusters that could not be identified based on marker gene expression and were excluded from further analysis (highlighted with dashed circles).
B. Marker gene expression across clusters. Colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression.
C. Gene Ontology (Biological Process) terms that are significantly enriched in marker genes for each cluster shown in Fig 1. The top most significant term per cluster was selected for inclusion in the plot. Size represents -log10 of the BH-adjusted p-value for enrichment of the term in the given gene set, and colour represents the number of DE genes associated with that term. Absent dots indicate no significant enrichment of that term in the given gene set.

1146

## Figure S2

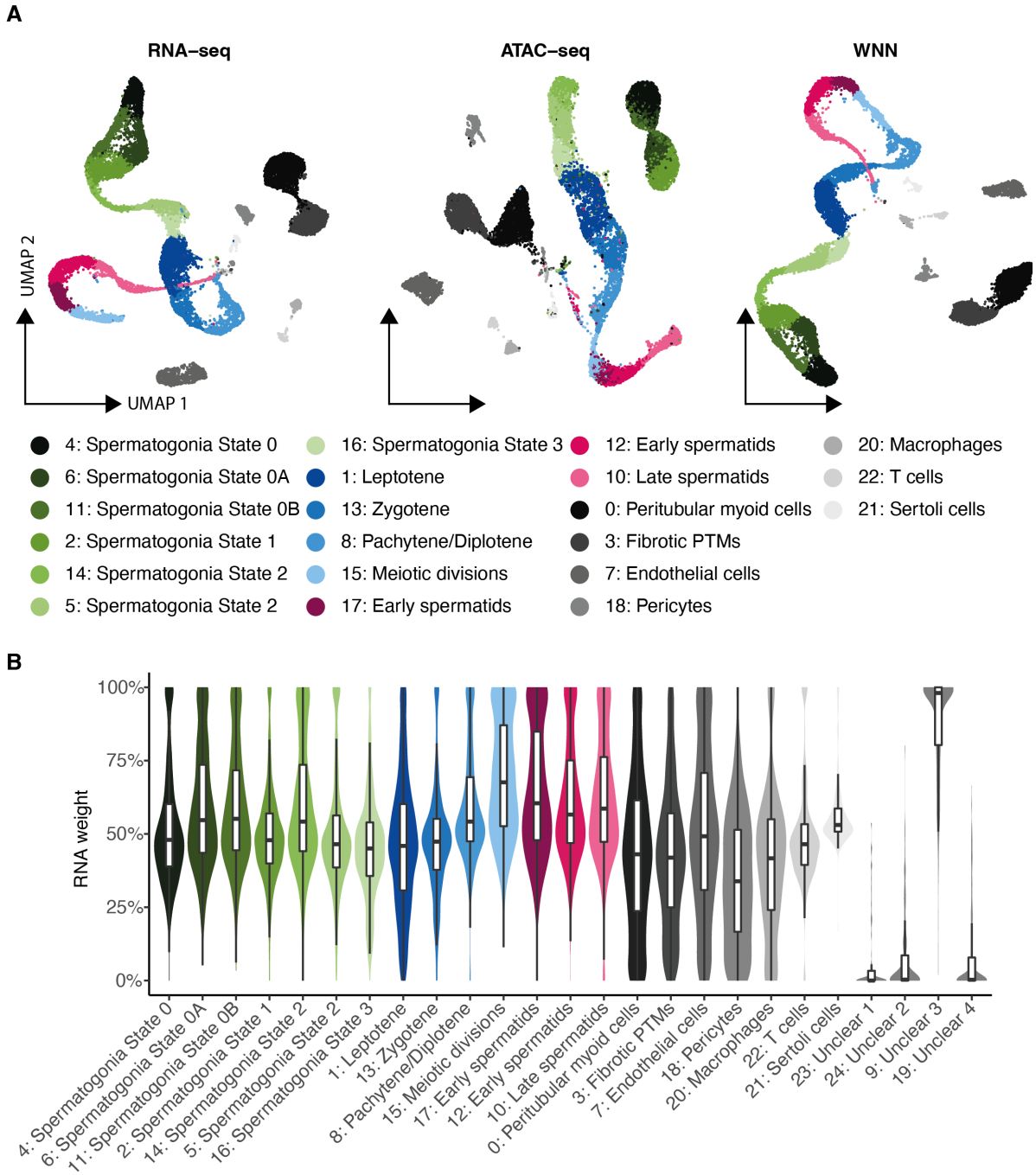



**Figure S2: RNA/ATAC-seq integration**

A. UMAPs based on RNA-seq, ATAC-seq, and the weighted nearest neighbour (WNN) graph produced by combining the two modalities using Seurat.

B. RNA/ATAC-seq weight per cluster. Most clusters have a 50:50 weighting of RNA and ATAC-seq; the 'Unclear' clusters are skewed towards 100% weighting of one modality.

**Figure S3**

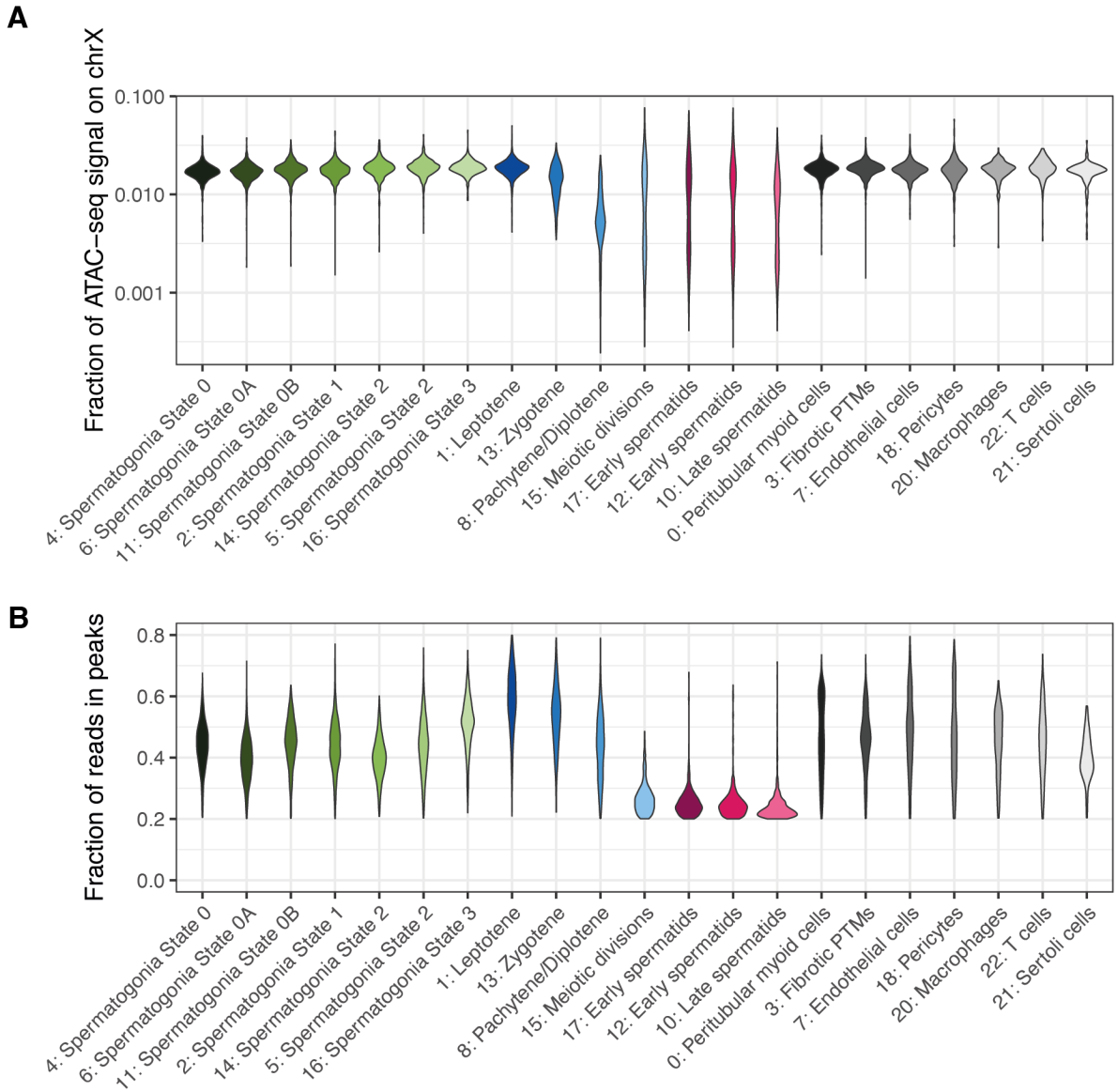



**Figure S3: Characteristics of ATAC-seq signal across clusters**

A. The fraction of all ATAC-seq reads in peaks coming from chrX is consistent throughout spermatogonia and in somatic cells, but drops in zygotene and pachytene/diplotene, and appears bimodal in meiotic divisions and spermatids, indicative of haploid cells containing either the X or Y chromosome.

B. The fraction of reads per cell which fall inside peaks drops sharply at the meiotic divisions stage and in spermatids. Cells with less than 20% of reads in peaks were filtered out during pre-processing as these are likely to be low quality cells.

## Figure S4

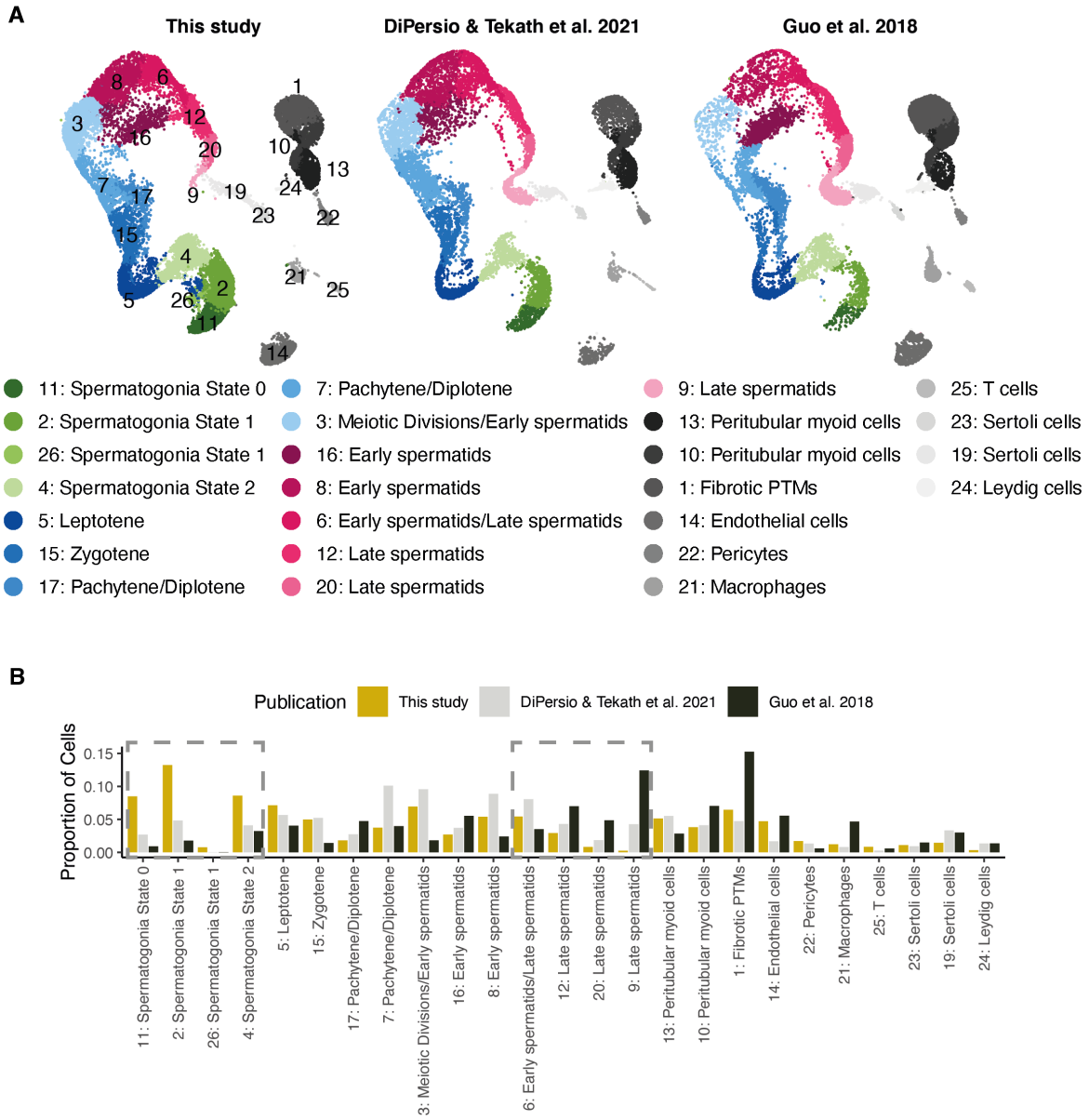



**Figure S4: Comparison of snRNA-seq and scRNA-seq.**

A. UMAPs of clusters obtained from snRNA-seq (this study) and published scRNA-seq from Di Persio & Tekath et al. 2021 and Guo et al. 2018[20,21].

B. Proportion of different cell types captured across the three datasets indicating an enrichment of spermatogonia and a depletion of spermatids in snRNA-seq compared to scRNA-seq datasets.

**Figure S5**

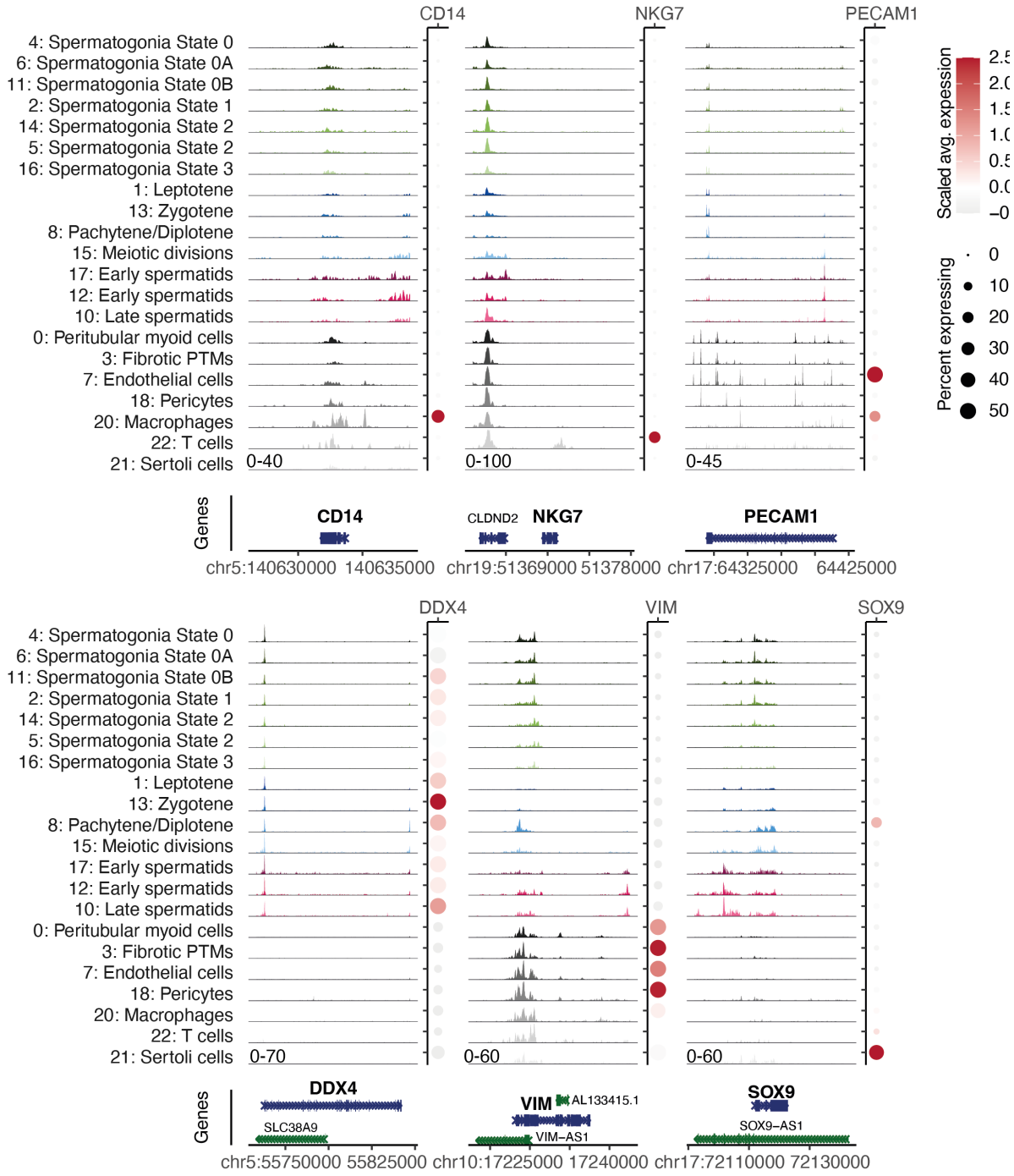



**Figure S5: Further examples of dynamic chromatin accessibility at cluster-specific marker genes.**

Each sub-panel shows cluster-specific pseudobulk chromatin accessibility tracks on the left. Expression of the marker gene of interest in the cluster shown in the dot plots on the right, where colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression. Marker genes shown are *CD14* for macrophages, *NKG7* for T cells, *PECAM1* for endothelial cells, *DDX4* for germline cells, *VIM* for somatic cells, and *SOX9* for Sertoli cells.

**Figure S6**

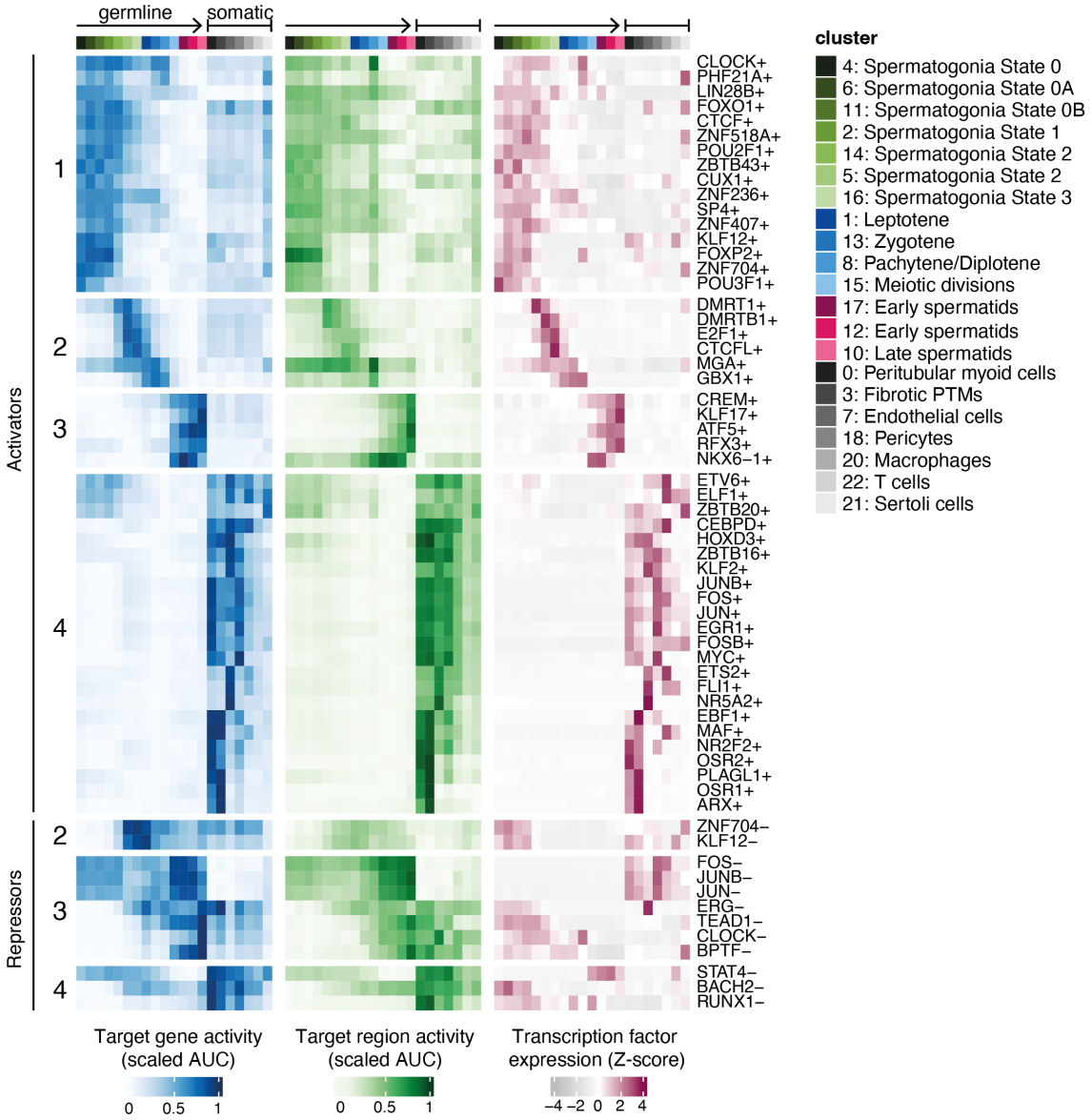



**Figure S6: Heatmap of eRegulon activity across clusters.**
62 high quality eRegulons which show dynamic activity across cell types were identified. TF name and direction of regulation are labelled on the right. Top, activator eRegulons. Bottom, repressor eRegulons. Left-to-right, panels represent eRegulon target gene expression enrichment per cluster (blue, scaled average AUC), eRegulon target region accessibility enrichment per cluster (green, scaled average AUC), and TF expression per cluster (magenta, Z-score). There are 4 distinct groups of eRegulons active in spermatogonia (1), differentiating spermatogonia and spermatocytes (2), spermatids (3) and somatic cell types (4).

1157 **Figure S7**

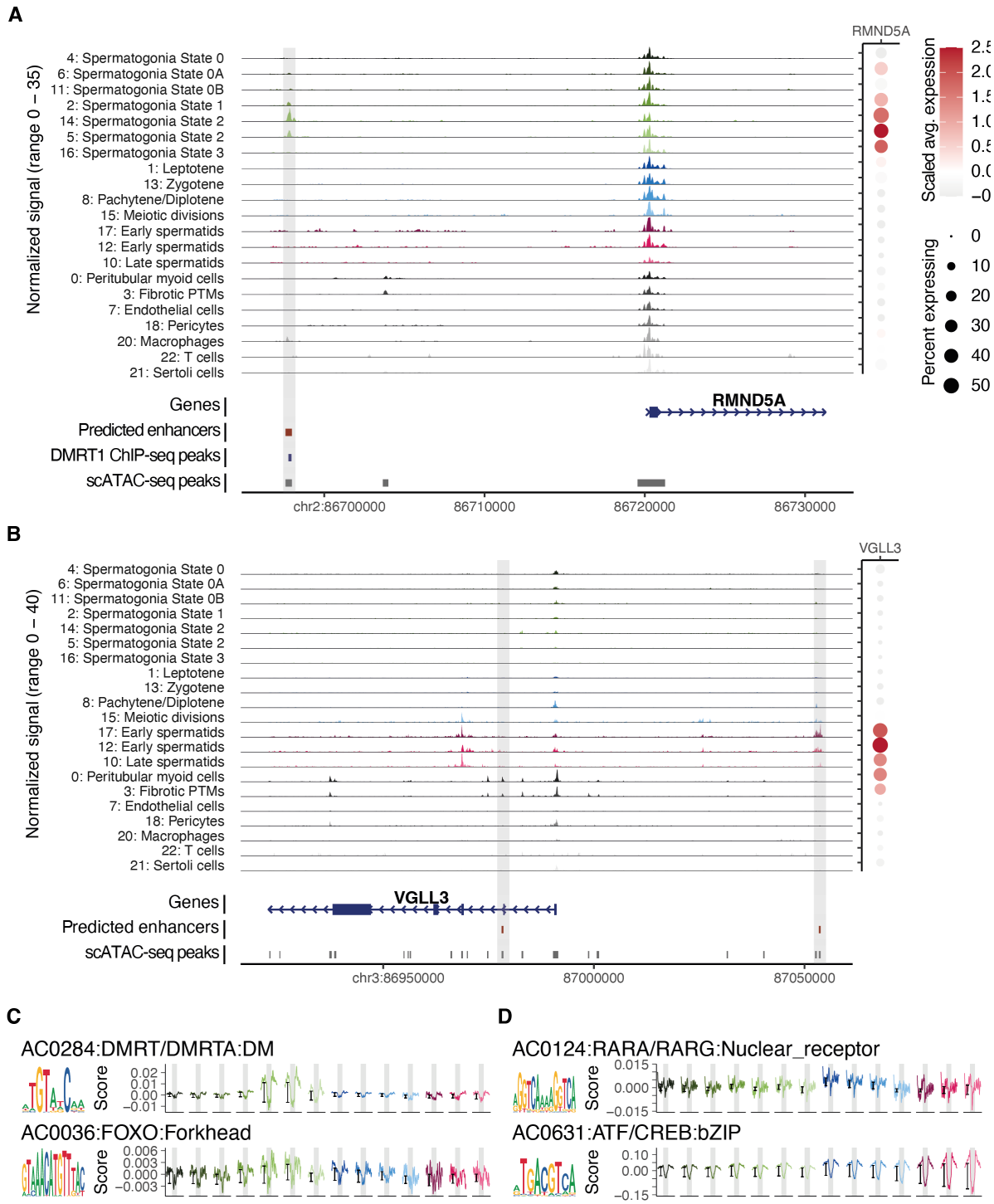



**Figure S7: Further examples of SCENIC+ predicted enhancers and target genes.**
A. SCENIC+ identifies an enhancer (red bar within grey highlight) upstream of *RMND5A* which is regulated by DMRT1, DMRTB1, and FOXO1. Left, cluster-specific pseudobulk chromatin accessibility tracks; right, *RMND5A* expression shown as a dotplot where colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression. Bottom, gene annotations, predicted enhancers (red), and scATAC-seq peaks (grey). A DMRT1 ChIP-seq peak overlapping the predicted enhancer is shown with a blue bar.

B. SCENIC+ identifies two enhancers for *VGLL3* (red bars within grey highlights). The lefthand enhancer is predicted to be regulated by EBF1 and NR2F2 and the righthand enhancer by ATF5 and CREM. Left, cluster-specific pseudobulk chromatin accessibility tracks; right, *VGLL3* expression shown as a dotplot where colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression. Bottom, gene annotations, predicted enhancers (red), and scATAC-seq peaks (grey).
C. TF footprinting for motifs bound by DMRT1/DMRTB1 (top) and FOXO1 (bottom), showing highest activity in differentiating spermatogonia.
D. TF footprinting for motifs bound by NR2F2 (top), showing little activity in germline cells, and by ATF5/CREM (bottom), showing highest activity in late spermatids.

**Figure S8**

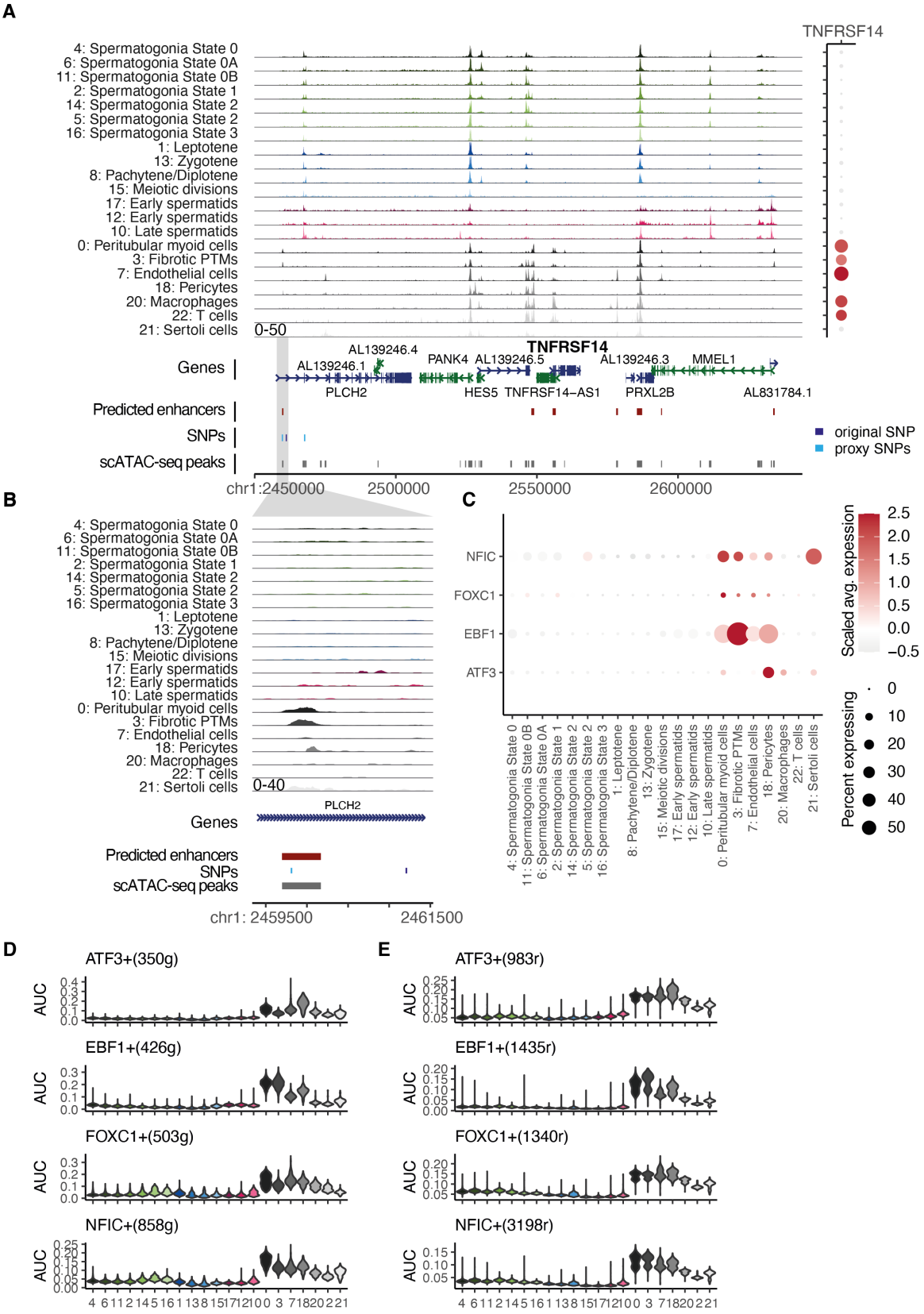



**Figure S8: SCENIC+ results identify *TNFRSF14* as a candidate target gene for a GWAS SNPs associated with infertility.**

A. *TNFRSF14* locus. Left, cluster-specific pseudobulk chromatin accessibility tracks. Right, *TNFRSF14* expression per cluster as a dotplot; colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression. Bottom, gene annotations, SCENIC+ predicted distal enhancers for *TNFRSF14*, SNPs, and snATAC-seq peaks. The original SNP (dark blue) is rs2477686 identified as associated with male infertility by Hu et al. 2012; proxy SNPs in LD ($r2 > 0.8$) with this SNP are shown in light blue.
B. Top, close up of part of the region in A. A proxy SNP of rs2477686 overlaps an snATAC-seq peak identified by SCENIC+ as a putative enhancer for *TNFRSF14*, regulated by ATF3, EBF1, FOXC1, and NFIC.
C. Expression of *ATF3*, *EBF1*, *FOXC1*, and *NFIC* as a dotplot; colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression.
D, E. eRegulon activity for ATF3, EBF1, FOXC1, and NFIC across clusters, as measured by target gene expression enrichment (D) and target region chromatin accessibility enrichment (E).

**Figure S9**

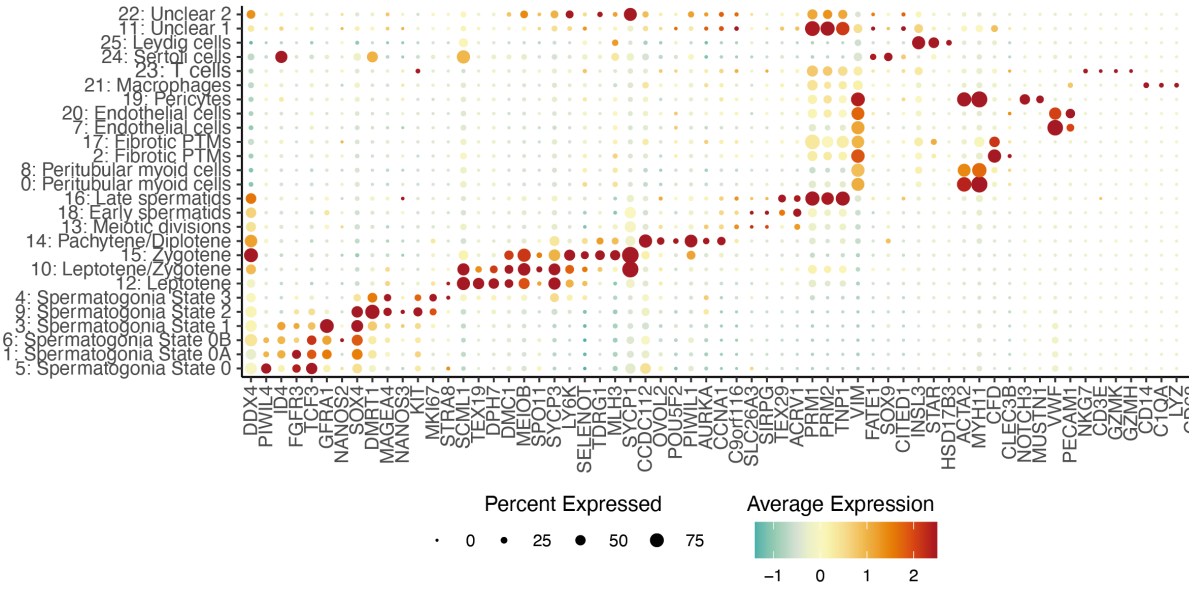



**Fig S9: Identification of clusters in the combined data set (normal + cryptozoospermia)**

Marker gene expression across clusters. Colour indicates scaled average expression and size indicates the % of cells in the cluster with detected expression.

**Figure S10**

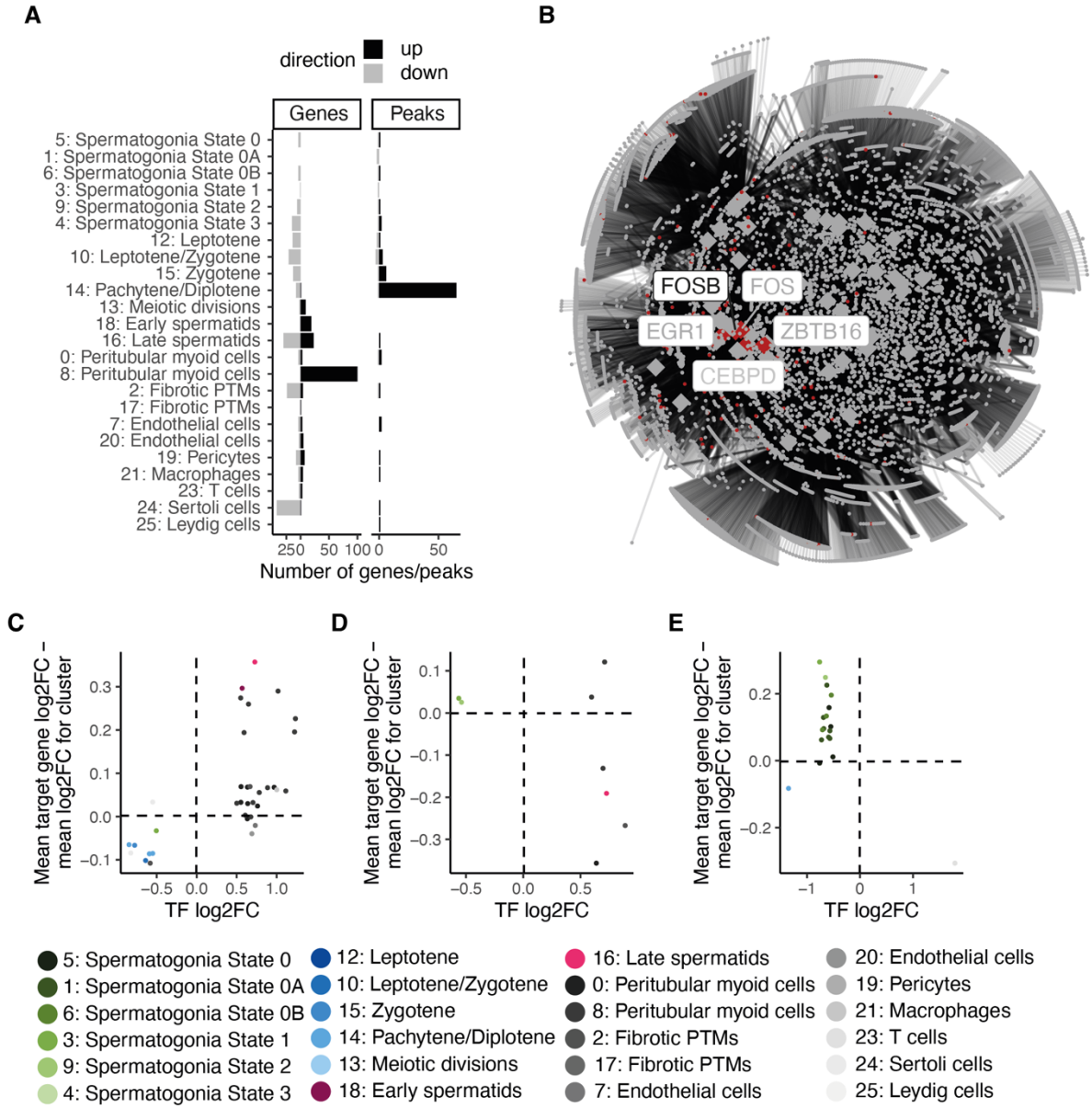



**Fig S10: Differential gene expression and chromatin accessibility in Crypto 1**

A. Numbers of differentially expressed genes (left) and differentially accessible peaks (right) in each cluster, comparing normal spermatogenesis and Crypto 1.

B. Differentially expressed genes between normal spermatogenesis and Crypto 1 shown in the context of the high-confidence gene regulatory network (Fig 2C). TFs are shown as large diamonds and all other genes as smaller circles. Genes that are significantly differentially expressed in any cluster are shown in red and all other genes in grey. Edges represent a regulatory relationship between a TF and a target gene inferred by SCENIC+. Differentially expressed TFs are labelled and coloured according to the cluster in which they have highest expression in the normal spermatogenesis samples.

C. The target genes of significantly differentially expressed SCENIC+ TFs show changes in gene expression consistent with SCENIC+ predictions. Activator TFs significantly differentially expressed in any cluster between normal and Crypto 1 ($p < 0.05$, absolute log2FC > 1) are shown. Each point represents a TF and its targets in a given cluster; the logFC of the TF is represented on the x axis and the difference in mean log2FC between

target genes and all genes is plotted on the y axis, as explained in Fig 4E. Clusters are coloured according to the key in A. Point size reflects the Fisher test p-value of the association between direction of change in expression for the TF and direction of change in expression of its targets.

D. As in C, for repressive TFs significantly differentially expressed in any cluster between normal and Crypto 1 ($p < 0.05$, absolute log2FC > 1).

E. As in C, for repressive TFs significantly differentially expressed in any cluster between normal and Crypto 2 ($p < 0.05$, absolute log2FC > 1).

1164

**Figure S11**

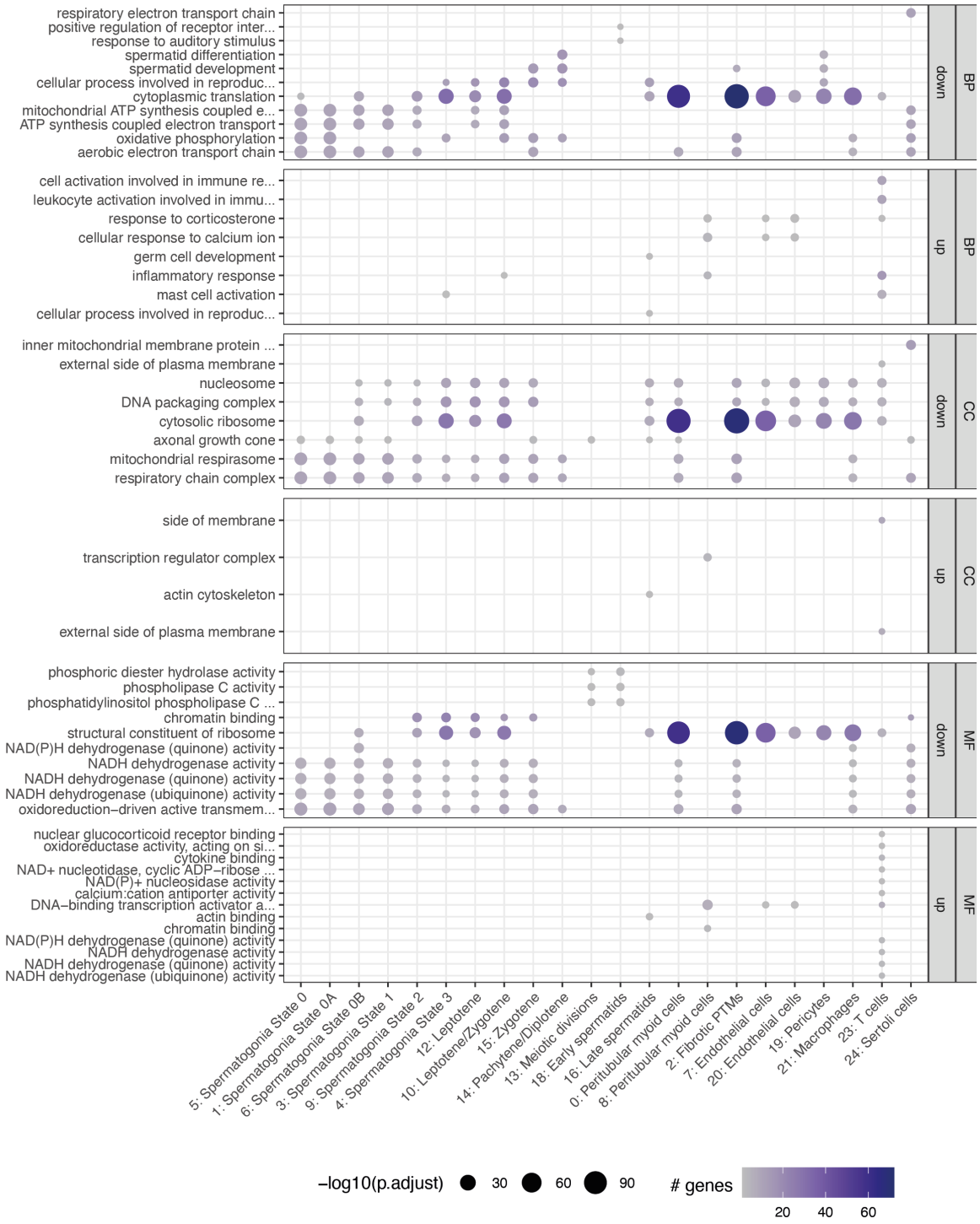



**Fig S11: Enriched GO terms for differentially expressed genes in Crypto 2**
The most significant GO term for each cluster is shown, considering up- and downregulated genes in each cluster separately. Size represents -log10 of the BH-adjusted p-value for enrichment of the term in the given gene set, and colour represents the number of DE genes associated with that term. Absent dots indicate no significant enrichment of that term in the given gene set. BP: biological process; CC: cellular component; MF: molecular function.

1166