Delineation of Meiotic Gene Expression in Male Mice



Daniel John Wells

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2020

Acknowledgements

This thesis would not have been possible without the support of many people, a few of whom I have space to thank below.

Individual

To my supervisors, Simon Myers and Jonathan Marchini, for taking me on as a student, for providing guidance and feedback, for having the patience to afford me the independence of carving my own path, and for your infectious enthusiasm for science.

To my collaborators and colleagues, Min Jung and Don Conrad; Emmanuelle Bitoun; Anjali Hinch, Gang Zhang, Peter Donnelly and Ben Davies; for generously sharing data and ideas.

To my first rotation supervisor Benjamin Schuster-Böckler for guiding me in my initial foray into genomics.

To my advisors and examiners, Ian Henderson, Gerton Lunter, Luke Jostins, Chris Holmes, and John Todd for their valuable advice and feedback on my work.

To Julian Knight and Isabel Schmidt, for supporting the Genomic Medicine and Statistics programme and community as well as their support and advice to me personally.

To my coursemates and officemates, for their comradeship and support.

To my parents, for their immeasurable love and support.

Institutional

To the Wellcome Trust for providing generous funding through the Genomic Medicine and Statistics programme (grant 109109/Z/15/Z).

To the Wellcome Center of Human Genetics and the Department of Statistics for providing office space and talks, and a supportive research environment.

To Linacre College, for providing accommodation, dining, and fitness facilities.

Abstract

Meiosis is a specialised cell division producing the haploid gametes required for sexual reproduction. A key function of this process is to generate genetic diversity through recombination and independent assortment of homologous chromosomes. Delineation of gene expression during this process has been challenging due to the heterogeneity of testis tissue and the lack of faithful *in vitro* models.

Here we jointly analyse a single-cell resolution transcriptomic dataset of over 20,000 cells from both wild type and mutant mice. We use dimensionality reduction methods to infer latent components of variation that represent transcriptional programmes, mutant specific pathological processes, and technical effects. This approach simultaneously soft clusters cells, infers corresponding groups of co-expressed marker genes, and imputes sparse, noisy gene expression. We were also able to infer, *de novo*, transcription factor binding motifs for each component, revealing a general switch at the meiotic divisions. We facilitate access to this resource by providing an interactive website testisatlas.ml.

The high-resolution delineation of gene expression during the spermatogenic programme provides high-resolution clues to the function of understudied genes. One such gene, *Zcwpw1*, is highly co-expressed with *Prdm9*, and has domains capable of recognising the unique combination of histone marks that PRDM9 deposits (H3K4me3 and H3K36me3). By using a human *in vitro* system of *Zcwpw1* co-transfection with either human or chimp *Prdm9*, we show that PRDM9 causes the recruitment of ZCWPW1 to its binding sites. This recruitment is stronger than for sites with H3K4me3 alone and is CpG dependent.

Male *Zcwpw1*$^{-/-}$ mice have completely normal double strand break positioning, but severe repair and synapsis defects leading to complete testicular azoospermia. Although PRDM9's effect of DSB positioning remains intact, PRDM9's effect of aiding synapsis appears to be abolished, with persistent DMC1 signal - most dramatically at the most strongly PRDM9 bound hotspots.

Contents

List of Figures **xi**

List of Abbreviations **xiii**

1 Background **1**
1.1 Aims . . . 2
1.2 Motivations . . . 2
1.2.1 Generating Maps of Gene Expression . . . 5
1.3 Measuring Gene Expression . . . 7
1.3.1 In Situ Hybridization . . . 8
1.3.2 RT-qPCR . . . 8
1.3.3 Northern Blot . . . 9
1.3.4 Microarrays . . . 9
1.3.5 Bulk RNA Sequencing . . . 9
1.3.6 Single Cell RNA Sequencing . . . 10
1.3.7 Spatial Sequencing . . . 13
1.3.8 Nascent RNA Sequencing . . . 13
1.3.9 Proteomics . . . 13
1.4 Gametogenesis . . . 14
1.4.1 Testis Anatomy . . . 14
1.4.2 Spermatocytogenesis . . . 16
1.4.3 Prophase I of Meiosis I . . . 17
1.4.4 Post-Pachytene Processes . . . 26
1.5 Prior Work . . . 28
1.6 Statistical Methods . . . 29
1.6.1 Dimensionality Reduction . . . 30
1.6.2 Sparse Decomposition of Arrays . . . 33

2 Single Cell Transcriptomics **37**
2.1 Introduction . . . 38
2.2 Experimental Data Generation . . . 39
2.3 Data Processing and QC . . . 39

2.4 SDA . . . 40
2.5 Low Dimensional Visualisation . . . 45
2.6 Imputation . . . 51
2.7 Mutant Genotypes . . . 54
2.8 Potential Artefactual Components . . . 59
2.9 Components . . . 60
2.9.1 Telocytes . . . 62
2.9.2 Leydig Cells . . . 62
2.9.3 Sertoli Cells . . . 63
2.9.4 Macrophages . . . 63
2.9.5 Lymphocytes . . . 64
2.9.6 Peritubular Myoid . . . 64
2.9.7 Spermatogonia . . . 64
2.9.8 (pre)Leptotene & Zygotene . . . 65
2.9.9 Pachytene . . . 67
2.9.10 Hormad1 Sex Chromosome Activation . . . 71
2.9.11 Cul4a, Mlh3, and Cnp Mutants . . . 73
2.9.12 Meiotic Divisions . . . 75
2.9.13 Round Spermatid . . . 76
2.9.14 Spermiogenesis . . . 81
2.10 Motif Inference . . . 83
2.10.1 Enrichment of Known Transcription Factor Targets . . . 83
2.10.2 *De-novo* inference from promoter sequences . . . 84
2.10.3 Broad-scale patterns of TFBS occurrence . . . 85
2.10.4 Novel Motifs . . . 88
2.11 Interactive Web App . . . 88
2.12 Detailed Methods . . . 90
2.12.1 Preprocessing of Drop-seq data . . . 90
2.12.2 Software . . . 90
2.12.3 X Chromosome Non Sharing Simulation . . . 91
2.12.4 Imputation Validation . . . 91
2.12.5 Transcription Factor Motif Analysis . . . 92
2.12.6 Data availability . . . 94

3 Zcwpw1 95
3.1 Introduction . . . 95
3.2 ZCWPW1 co-evolves with PRDM9 . . . 97
3.3 ZCWPW1 localisation and KO phenotypes . . . 99
3.4 ZCWPW1 recruitment by PRDM9 . . . 104

3.5 DSBs positioning and repair . . . 109
3.6 ZCWPW1 binds CpG dinucleotides . . . 113
3.7 Discussion . . . 116
3.8 Methods . . . 120
3.8.1 Orthologue Alignment . . . 120
3.8.2 Domain Search . . . 122
3.8.3 Mice and genotyping . . . 123
3.8.4 Fertility measurements . . . 124
3.8.5 Immunostaining of spermatocytes . . . 124
3.8.6 FISH . . . 125
3.8.7 ZCWPW1 antibody production . . . 125
3.8.8 Cell line, transfection, and IF staining . . . 126
3.8.9 Immunoprecipitation and Western Blot detection . . . 126
3.8.10 ChIP-seq . . . 127
3.8.11 Data and Code availability . . . 132

4 Perspectives **133**
4.1 Other Single Cell RNAseq studies . . . 133
4.2 Observational Extensions . . . 139
4.3 Other Zcwpw1 Papers & Extensions . . . 140

Appendices

A Supplementary Figures **145**

References **161**

List of Figures

1.1 Testis PCA (MCA) . . . 4
1.2 scRNAseq Publications . . . 11
1.3 Testis Anatomy . . . 15
1.4 Synaptonemal Complex . . . 18
1.5 Molecular Recombination . . . 20
1.6 Spermatids . . . 27
1.7 SDA Matrix Factorisation . . . 33
1.8 SDA Model . . . 34

2.1 A Single Cell Component . . . 41
2.2 SDA Convergence . . . 43
2.3 10k vs 1k SDA Iterations . . . 44
2.4 t-SNE random seed stability . . . 46
2.5 Marker Genes . . . 47
2.6 SDA . . . 48
2.7 t-SNE of Components . . . 49
2.8 Overlapping SDA Components . . . 50
2.9 SDA Imputation . . . 52
2.10 Imputation Supplement . . . 53
2.11 Wild Type vs Mixed SDA . . . 56
2.12 Procrustean Rotation Matrix . . . 57
2.13 Component 31 WT vs Mixed . . . 58
2.14 Non Meiotic Components . . . 60
2.15 SDA Cell Scores . . . 61
2.16 Meiotic Sex Chromosome Inactivation . . . 69
2.17 MSCI Supplement . . . 70
2.18 Hormad1 KO Component . . . 72
2.19 Misc. Components . . . 73
2.20 Experimental groups in t-SNE space . . . 75
2.21 Novel Immunofluorescence Markers . . . 79
2.22 Known Transcription Factor Targets Enrichment . . . 84
2.23 Denovo Motifs . . . 85

2.24 Motifs Heatmap 86
2.25 Interactive Web Application 89

3.1 Zcwpw1 Domains and Conservation 98
3.2 Zcwpw1 Expression 100
3.3 Zcwpw1 KO Phenotypes 102
3.4 DMC1 Counts 103
3.5 Zcwpw1 Allele Specificity 105
3.6 Dual mark recruitment 108
3.7 DMC1 SSDS ChIPseq 111
3.8 ZCWPW1 binds CpG rich sequences 115

A.1 *Zcwpw1* Tissue and Isoform expression 145
A.2 Telomeric Localisation 146
A.3 Tangled Chromosome Phenotype 146
A.4 RAD51 Counts 147
A.5 Zcwpw1 alone vs Co-transfection 148
A.6 Transfection Efficiency 149
A.7 Zcwpw1 Heatmaps 150
A.8 DMC1 prediction 151
A.9 Histone Correlation from peaks 152
A.10 Histone Correlation from windows 153
A.11 Relative H3K4me3 Strength 154
A.12 Hotspot Sharing 155
A.13 DMC1 positions 156
A.14 HOP2 KO Phenocopying 156
A.15 DMC1 ratio regression 157
A.16 ChIPseq peaks proportions 158
A.17 CpG counts for various repeats 159

List of Abbreviations

AUC . . . . . Area Under the Curve

ChIP . . . . . Chromatin Immuno-precipitation

CI . . . . . Confidence Interval

CpG . . . . . Cytosine phosphodiester Guanine

DSB . . . . . Double Strand Break

dHJ . . . . . Double Holliday Junction

DGE . . . . . Digital Gene Expression (count matrix)

E10.5 . . . . . Embryonic Day 10.5

FACS . . . . . Fluorescence Activated Cell Sorting

FET . . . . . Fisher's Exact Test

FISH . . . . . Fluorescent In Situ Hybridisation

GAM . . . . . Generalised Additive Model

GO . . . . . Gene Ontology

GWAS . . . . . Genome Wide Association Study

H3K4me3 . . . Histone 3 Lysine 4 tri-methylation

Het . . . . . Heterozygous

Hom . . . . . Homozygous

ICA . . . . . Independent Component Analysis

KO . . . . . Knock Out

MCA . . . . . Mouse Cell Atlas

MSCI . . . . . Meiotic Sex Chromosome Inactivation

mRNA . . . . messenger Ribose Nucleic Acid

NNMF/NMF Non-Negative Matrix Factorisation

NGS . . . . . Next Generation Sequencing

OR . . . . . Odds Ratio

PCA . . . . . Principal Component Analysis

PCR . . . . . Polymerase Chain Reaction

PIP . . . . . Posterior Inclusion Probability

PND5 . . . . . Post Natal day 5

QC . . . . . Quality Control

RT . . . . . Reverse Transcription

SDA . . . . . Sparse Decomposition of Arrays

scRNAseq . . Single Cell RNA sequencing

SNP . . . . . Single Nucleotide Polymorphism

SPCI . . . . . Primary Spermatocyte

SPCII . . . . . Secondary Spermatocyte

SPG . . . . . Spermatogonia

SPD . . . . . Spermatid

SSDS . . . . . Single Stranded DNA Sequencing

TF . . . . . Transcription Factor

TFBS . . . . . Transcription Factor Binding Site

TSS . . . . . Transcription Start Site

t-SNE . . . . . t-distributed Stochastic Neighbourhood Embedding

UMAP . . . . Uniform Manifold Projection

UMI . . . . . Unique Molecular Identifier

UTR . . . . . Untranslated Region

WT . . . . . Wild Type

Es müsse eine Art der Kerntheilung geben, durch welche die im Mutterkern enthaltenen Ahnenplasmen dergestalt auf die Tochterkerne vertheilt werden, dass jedem Tochterkern nur die halbe Zahl derselben zukomme.

There must be a form of nuclear division in which the ancestral germ plasms contained in the nucleus are distributed to the daughter nuclei in such a way that each of them receives only half the number contained in the original nucleus.

— Weismann, 1887

1 Background

Contents

1.1 Aims . . . 2
1.2 Motivations . . . 2
1.2.1 Generating Maps of Gene Expression . . . 5
1.3 Measuring Gene Expression . . . 7
1.3.1 In Situ Hybridization . . . 8
1.3.2 RT-qPCR . . . 8
1.3.3 Northern Blot . . . 9
1.3.4 Microarrays . . . 9
1.3.5 Bulk RNA Sequencing . . . 9
1.3.6 Single Cell RNA Sequencing . . . 10
1.3.7 Spatial Sequencing . . . 13
1.3.8 Nascent RNA Sequencing . . . 13
1.3.9 Proteomics . . . 13
1.4 Gametogenesis . . . 14
1.4.1 Testis Anatomy . . . 14
1.4.2 Spermatocytogenesis . . . 16
1.4.3 Prophase I of Meiosis I . . . 17
1.4.4 Post-Pachytene Processes . . . 26
1.5 Prior Work . . . 28
1.6 Statistical Methods . . . 29
1.6.1 Dimensionality Reduction . . . 30
1.6.2 Sparse Decomposition of Arrays . . . 33

1.1 Aims

Meiosis is a type of cell division in which the number of chromosomes is reduced by half to produce haploid gametes (sperm and egg cells) (Ohkura, 2015). The primary aim of this research was to delineate the gene expression programme of spermatogenesis (during which meiosis occurs). Specifically, to identify which genes are expressed, in which cells, their degree of expression, and relative timing during spermatogenesis. With this atlas of expression data, we additionally aimed to infer the regulatory network controlling these expression patterns, as well as to predict protein function and in doing so identify targets for further characterisation. Finally, we aimed to carry out such characterisation for at least one of these targets.

1.2 Motivations

Reproduction is a fundamental property of all known life. Of the three domains of life Eukaryotes/Eukarya are able to reproduce sexually, and almost all do. The dominance of this trait has been estimated at roughly 99.9% (White, 1978). Although many Eukaryotes are able to reproduce asexually, those which do so exclusively are typically confined to the youngest branches of life, an often cited exception being the Bdelloid rotifers - although there is now evidence of rare sexual reproduction in this species (Debortoli et al., 2016; Laine et al., 2020). This shared characteristic was likely present in the common ancestor of this empire and hence synapomorphic (Bernstein et al., 2013). The sexual life cycle consists of an alternating halving of the genetic material/ploidy (by meiosis) to form haploid cells, followed by karyogamy (nuclear fusion). Meiosis (from Ancient Greek, meíōsis, "a lessening") is a specialised form of cell division, in which genetic material is exchanged between the parental genomes (recombination). Recombination and independent assortment of homologous chromosomes generate genetic diversity and hence enable natural selection and evolution. Recombination also ensures balanced segregation at the reductional division which follows (no naturally occurring

eukaryotes have a single chromosome - although synthetic yeast has been engineered - Y. Shao et al., 2018).

The mechanisms of meiosis have important consequences for the evolution, fertility, and speciation of sexually reproducing organisms (B. Davies et al., 2016; Hassold et al., 2007).

As the process of gamete generation, errors in meiosis can cause genetic disease in offspring. In some sense the most extreme version is complete infertility, affecting around 10% of humans (with many being of unexplained cause) (Datta et al., 2016; Hamada et al., 2011). Other errors can be compatible with fertilisation but not full gestation resulting in miscarriage (affecting around 13% of pregnancies overall, but over 50% for mothers over the age of 45 years - Magnus et al., 2019). Others errors are compatible with birth but result in incurable disease such as some aneuploidies (Hassold et al., 2007) and microdeletion syndromes caused by non-allelic homologous recombination (S. Myers et al., 2008).

A better understanding of the mechanisms of spermatogenesis and meiosis could aid in the prevention of such diseases. Additionally, determining the factors required for *in vitro* spermatogenesis could enable the treatment of male infertility in addition to facilitating further research (Q. Zhou et al., 2016).

While enabling the production of healthy offspring is one aim of reproductive medicine, *preventing* fertilisation (usually but not always, temporarily) is also extremely important for the health of both mother and child (Cleland et al., 2012). Hence the development of contraceptives is another key goal of reproductive science, and one which a clearer delineation of testis specific genes could aid (N. Schultz et al., 2003).

As with many scientific endeavours, curiosity is also a major driver of this research. The testis is a most unusual tissue especially at the transcriptional level. The testis has the highest number of tissue enriched genes out of the 32 tissues profiled in the Human Protein Atlas (> five-fold higher mRNA levels in one particular tissue as compared to any other), with 1,057 enriched genes out of a total 2,489 across the 32 tissues (Djureinovic et al., 2014; Mele et al., 2015; Uhlén

et al., 2015; Uhlén et al., 2016; The Human Protein Atlas, 2019). This is over twice as many as the second highest tissue the cerebral cortex which is perhaps considered the most complex tissue functionally, and with which the testis shares a surprising similarity (J. H. Guo et al., 2005; Djureinovic et al., 2014; Uhlén et al., 2015). More recent studies using single-cell sequencing have underscored the distinct transcriptional state of the testis (Han et al., 2018 and figure 1.1). The testis is also one of a few immune privileged sites in the human body (Fijak et al., 2006).

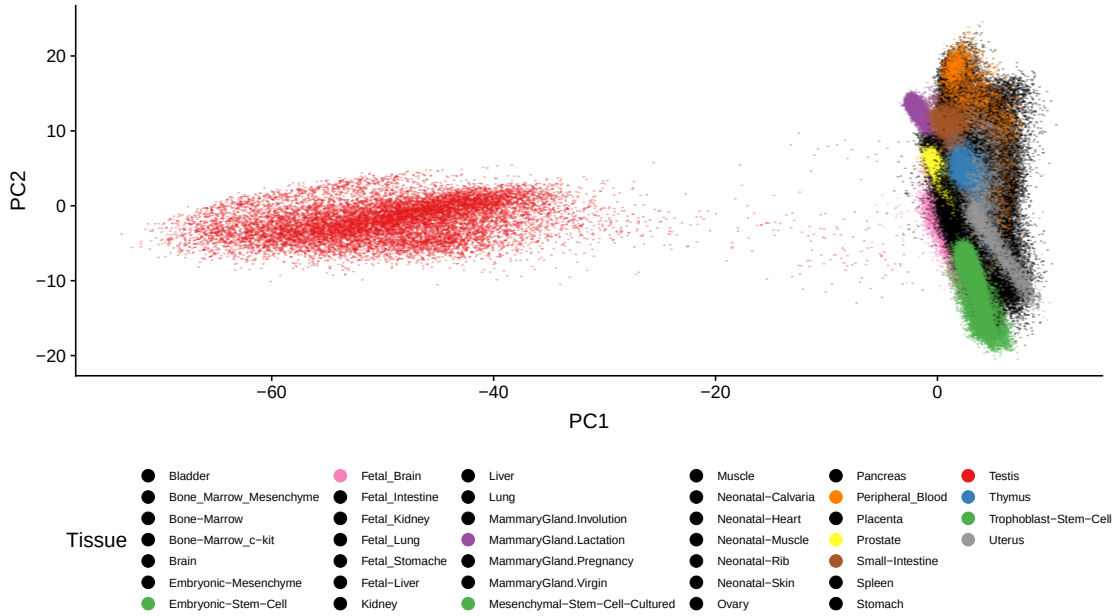


Figure 1.1: PCA of single cell RNAseq of 50 different tissues and cell types from over 400,000 cells from the mouse. Data from Han et al., 2018

The previous two properties have lead to the use of testis expressed genes as targets for cancer immunotherapy (cancer testis antigens) as when re-expressed in cancers they provide a distinguishing biomarker (Whitehurst, 2014).

Chromatin undergoes massive remodelling during spermatogenesis ultimately resulting in most (90/99% in human/mouse) of the histones being replaced by crosslinked protamines, leading to the cessation of transcription (Rathke et al., 2014). This necessitates long term mRNA storage (>5 days) before translation (Kleene, 2013).

Sperm are a highly specialised cell, despite being a single cell and unable to transcribe new mRNA they must perform many of the functions of the whole organism including motility, sensing the environment (chemotaxis), and defence from attack (from the female immune system) (Kaupp et al., 2008; L. A. Thompson et al., 1992). Indeed they are able to survive and even fertilise after more than three days in the female reproductive tract (Gould et al., 1984; Wilcox et al., 1995).

Meiosis itself of course is also unique, with paternal and maternal homologous chromosomes being relatively unaware of each other in somatic cells, until pairing in meiosis in the germline. In addition, it has been repeatedly observed that the majority (~80%) of inherited single nucleotide mutations first arise in the testis (Kong et al., 2012; Michaelson et al., 2012; Goldmann et al., 2016; Rahbari et al., 2016; Jónsson et al., 2017).

1.2.1 Generating Maps of Gene Expression

Much of science can be characterised as the creation of maps. The discovery, description and cataloguing of various types of elements, which is often dismissed as 'merely descriptive', is a fundamental bedrock of many sciences (Grimaldi et al., 2007). Maps are often useful in two senses, they serve as a reference of what's already known, but they also arrange current knowledge in such a way as to provide a prediction of the unknown. For example Mendeleev's periodic table both catalogued the known elements but left gaps for new discoveries. Equally with the Standard Model in physics. Darwin and others made extensive catalogues of past and present diversity of life, eventually resulting in the concept of evolution by natural selection. Cataloguing the ratios of offspring phenotypes lead to Mendel's laws.

Our maps of protein function are woefully incomplete, with more than 3,000 (15%) of human genes not even having a biological process annotation (and a similar % in the more experimentally amenable organisms *S. cerevisiae* and *S. pombe*) (V. Wood et al., 2019). Even genes with at least one annotation or paper are often incompletely or even incorrectly characterised. Significant bias exists against initiating research into a gene of unknown function (Edwards et al., 2011;

Stoeger et al., 2018; Haynes et al., 2018). This can be partially attributed to risk averse incentives in scientific funding and evaluation. In addition, for many proteins of unknown function there are limited tools available such as antibodies, making even simple characterisation difficult (for example immunofluorescence or immuno-precipitation based assays).

Genome sequencing projects have provided single base resolution maps of many organisms' genomes (thousands of eukaryotes - *Genome List - Genome - NCBI* 2019). This relatively unbiased approach has uncovered many genes of unknown function and has accelerated their characterisation. For example, by enabling the inference of molecular function from homology with known domains, and by facilitating sequence based detection and cloning. Ultimately much of what we care about is at the level of a phenotype such as infertility, and GWA studies provide one way to potentially associate genes with such phenotypes (Gajbhiye et al., 2018). However, even *if* a causal gene can be confidently identified for an associated SNP, because the phenotypes are usually broad, this often tells us little about the molecular function of a gene, necessitating substantial follow up research which is frequently not practical or undertaken (Visscher et al., 2017; Gallagher et al., 2018; Struck et al., 2018).

In a reverse genetics approach, specific mutations are created and the phenotypic results are investigated. The International Mouse Phenotyping Consortium aims to generate and broadly phenotype knockout mice for every protein coding gene, with a current catalogue of over 5,000 null mice models ("The Knockout Mouse Project" 2004; Meehan et al., 2017; Birling et al., 2019).

Other atlases such as the Human Protein Atlas and GTEx instead start from the first step in a gene's function, providing transcriptomic data enabling association of co-expressed genes and association of tissue specific genes with the function of those tissues. However, bulk tissues are too heterogeneous and large scale to provide precise insights into molecular function. Through a similar argument, proteins with co-varying abundances and protein-protein interactions can also implicate functions.

A protein's structure is often more conserved than its sequence and can give direct mechanistic insight into its function (Illergård et al., 2009; Sousounis et al., 2012). Despite being less able to take advantage of advances in high throughput next generation sequencing technology, over 144,000 structures are held by the Protein Data Bank, many deposited by structural genomics programs, but still many protein families lack a single structure (wwPDB consortium, 2019; Grabowski et al., 2016; Khafizov et al., 2014).

Recently, droplet based single cell RNAseq technologies have enabled higher resolution high-throughput characterisation of gene co-transcription. A number of mapping projects have been launched including the Human Cell Atlas and Mouse Cell Atlas as well as other tissue specific maps including the one we present here, helping to provide much higher resolution clues for the function of unstudied genes (Regev et al., 2017; Regev et al., 2018; Han et al., 2018; The Tabula Muris Consortium, 2018). We will show that this allows us to more reliably associate particular groups of genes with specific functions, at specific times, in meiosis.

1.3 Measuring Gene Expression

Expression is the process through which information contained within the genome manifests as a phenotype. This involves generation of a functional biomolecule, such as proteins but sometimes RNA (most importantly the ribosome). The first step in expression is the generation of messenger(m) RNA and there are a number of different technologies available to quantify the amount of mRNA - each providing different types and amounts of information.

Physical limits make it challenging to sensitively and accurately quantify gene expression. mRNA is <5% of total RNA in a cell, with the rest comprising of ribosomal and transfer RNA (Warner, 1999). Total RNA content per eukaryotic cell is approximately 20pg, with <1pg of mRNA, corresponding to roughly <500,000 mRNA molecules (Roozemond, 1976; Uemura, 1980; Tang et al., 2011). Most genes have mRNA copy numbers of 1-100 per cell, up to a maximum of perhaps 1,000 (Marguerat et al., 2012; Macaulay et al., 2014). At the level of individual

cells, mRNA synthesis is "bursty" introducing further biological heterogeneity in expression measurements (Raj et al., 2006; Chubb et al., 2006).

1.3.1 In Situ Hybridization

In Situ Hybridisation, first described by Gall et al., 1969, uses a probe (either RNA or DNA) which is incubated with fixed cells and binds to a target mRNA by complementary binding and can then be visualised. This has the advantage of being single cell resolution, and even allowing subcellular localisation to be observed.

Visualisation of the probe initially used radioactive labelling which was low resolution, expensive, unstable, and dangerous. However, this was overcome by using biotinylated probes coupled with secondary detection (Singer et al., 1982) and later directly fluorescent probes (FISH) (Kislauskis et al., 1993).

This sensitivity of this technique has been further increased by using multiple probes per target RNA, enabling the detection of individual mRNA molecules (single molecule smFISH - Femino et al., 1998; Raj et al., 2008). In addition, multiplexed detection schemes using multiple cycles of sequential hybridization and imaging such as SeqFISH(+) (Lubeck et al., 2014; Eng et al., 2017; Eng et al., 2019) and MERFISH (K. H. Chen et al., 2015; Moffitt, J. Hao, G. Wang, et al., 2016; Moffitt, J. Hao, Bambah-Mukku, et al., 2016; C. Xia, Babcock, et al., 2019) are able to measure many (up to 10,000) RNA targets. These methods have the advantage that amplification is not necessarily required and the high sensitivity allows single molecules to be detected. Furthermore, by measuring nuclear vs cytoplasmic localisation or by using probes to introns, mRNA dynamics ("RNA velocity") and isoform usage can also be measured (Shah et al., 2018; C. Xia, J. Fan, et al., 2019).

1.3.2 RT-qPCR

Reverse transcription quantitative real-time PCR measures PCR products during the amplification by using fluorescent detectors (Gibson et al., 1996; Heid et al., 1996; Chiang et al., 1996). This technique is able to detect a single mRNA molecule, and has become the gold standard in expression quantification. However, primers

must be designed for each gene and any spatial information is lost (Palmer et al., 2003; M. L. Wong et al., 2005).

Furthermore, due to the exponential amplification, small differences in reverse transcription, primer efficiency, or initial starting amounts and contamination can lead to inaccurate quantification. Therefore, efficiency and validated reference genes must be established by separate experiments (Bustin et al., 2009).

1.3.3 Northern Blot

The Northern blot is one of the earliest methods for measuring mRNA levels (Alwine et al., 1977). It consists of RNA extraction, separation by gel electrophoresis, followed by transfer by blotting onto a membrane and finally detection by complementary hybridisation with a radioactive/fluorescent probe. Northern blotting still remains a useful technique when information about the relative sizes of transcripts is required.

1.3.4 Microarrays

A microarray is typically a glass slide coated with cDNA oligomers. mRNA with polyA tails can be extracted from total RNA using oligo(dT) beads, reverse transcribed to cDNA, fluorescently labelled, and washed over the array (Schena et al., 1995; DeRisi et al., 1996; Schulze et al., 2001). The fluorescent intensity is a measure of the abundance of mRNA in the sample. This approach allowed whole transcriptome quantification and was wildly popular, with the EBI ArrayExpress database passing 100,000 hybridisations in 2008 (Rustici et al., 2008).

1.3.5 Bulk RNA Sequencing

The reducing cost rapidly led to the use of DNA sequencing technology to quantify the transcriptome (Mortazavi et al., 2008; Cloonan et al., 2008; Wilhelm et al., 2008; Lister et al., 2008; Nagalakshmi et al., 2008). mRNA is extracted as for microarrays, fragmented, and sequenced using short read NGS technology before being re-mapped to the genome. This has the advantage over microarrays that

the sequence of hybridisation probes does not need to be designed from an already known transcriptome reference, enabling, for example, novel transcripts and splice isoforms to be discovered. RNAseq provides a digital read out, in comparison to the analogue signal from microarrays which limited its replicability across platforms. In addition, the dynamic range is much greater than for microarrays in which the signal saturates when all probes are bound and at low concentrations non-binding is favoured (Zhao et al., 2014). RNAseq also enables detection of sequence variants and allele specific expression. For these reasons RNAseq has become the standard method for profiling the transcriptome (Stark et al., 2019).

As it is based on standard sequencing technology RNAseq, benefits from advances in next generation sequencing. For example, long read technology from PacBio enables full transcript (cDNA) sequencing - enabling the unambiguous identification of transcript isoforms which previously had to be inferred using the fragmentation approach of short read sequencing (Sharon et al., 2013; Cartolano et al., 2016). Oxford Nanopore sequencing infers the sequence of polynucleotides from the change in current as they pass through a nanopore embedded within a membrane across a voltage. In addition to being able to sequence long reads as PacBio (Oikonomopoulos et al., 2016), this technology was recently demonstrated to be able to sequence mRNA directly, bypassing the fragmentation, reverse transcription, PCR amplification and size selection steps, reducing opportunities for bias (Garalde et al., 2018; Workman et al., 2019). Furthermore Oxford Nanopore sequencing is able to quantify polyA lengths and detect RNA modifications such as N6-methyladenine.

1.3.6 Single Cell RNA Sequencing

RNA sequencing of single cells (scRNAseq) was first performed by Tang et al., 2009 using a modified amplification protocol previously used in single cell microarray studies, generating a dataset of five cells. By incorporating a cell-unique barcode during reverse transcription, cells can be multiplexed (pooled post mRNA capture and separated in silico) reducing cost and increasing the cell throughput to ~100 (Islam et al., 2012). Using automated liquid handling robots and 384 well plates

enabled the generation of a 4,468 cell dataset (Jaitin et al., 2014). Such scale is also possible using subnanolitre well arrays (Bose et al., 2015; H. C. Fan et al., 2015; Gierahn et al., 2017). The use of oil immersed nano-droplets to encapsulate individual cells with lysis buffer and polyT coated beads further increased throughput up to 44,808 cells (Macosko et al., 2015; Klein et al., 2015). Both microwell and nanodroplet isolation methods have now been used to generate datasets at an additional magnitude of scale with 400k, and 690k cells respectively (Saunders et al., 2018; Han et al., 2018). This exponential scaling of single cell RNAseq datasets appears to be continuing with one dataset including more than 2 million cells (Cao et al., 2019) and creation of datasets with tens of thousands of cells now being common, especially using the commercial Chromium platform from 10X genomics (figure 1.2, G. X. Y. Zheng et al., 2017; Svensson et al., 2018).

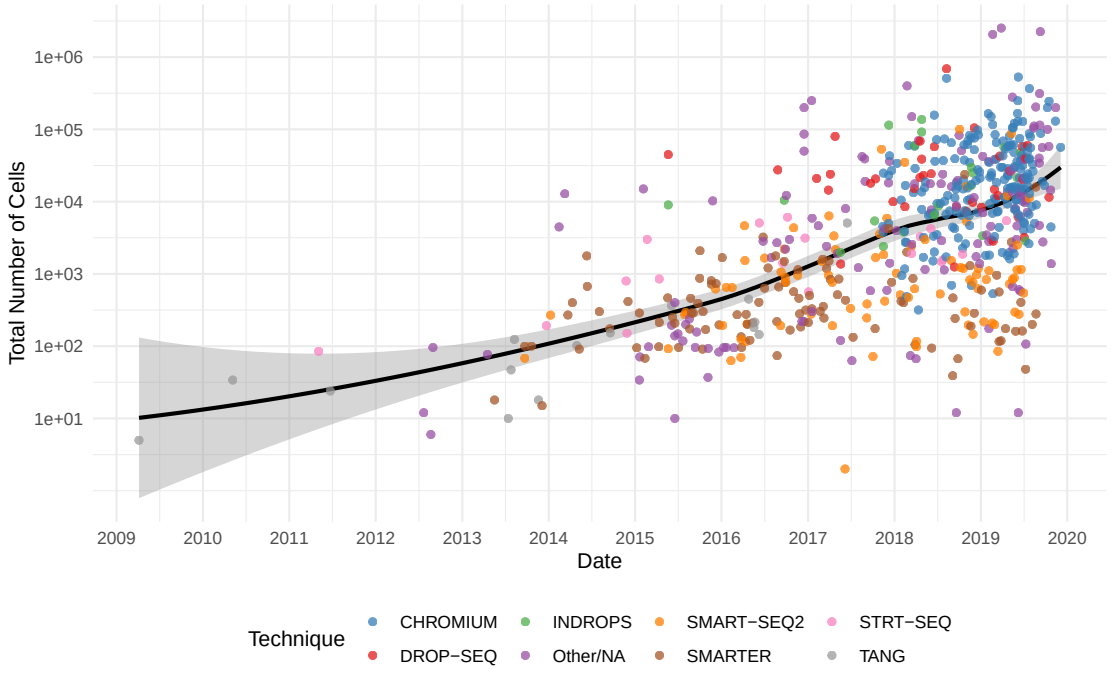


Figure 1.2: Exponential scaling of Single Cell RNAseq studies. Data from Svensson, Beltrame, et al., 2019

In scRNAseq, due to the low amount of starting material, many rounds of PCR amplification are required to generate sufficient material for sequencing. However this amplification step introduces significant bias, skewing the resulting levels (Best

et al., 2015). To alleviate this unique barcodes named UMIs (unique molecular identifier) can be included in each poly-dT probe, in addition to the unique barcode for each cell (1 per bead) (Kivioja et al., 2012; Islam et al., 2014). This enables each original mRNA molecule to be counted as a single molecule, in effect providing an absolute count of molecules rather than relative abundance as in standard bulk RNAseq. However, UMIs do not alleviate bias due to preferential sampling of molecules by RT or PCR primers (Kou et al., 2016). In addition, as the UMI and cellular barcode is on the 3' end of the transcript, after fragmentation only the fragments at the 3' end will be sequenced. This bias is in addition to the 3' bias from polyA selection (Wan et al., 2012; Lahens et al., 2014). This 3' bias precludes analysis of transcript isoforms as only the 3' end of the fragment is sequenced (as most transcripts share the 3' UTR, and the 5' read contains only the UMI and cell barcode). However, some isoform information is available due to the use of alternative 3' UTRs (and possibly by internal polyA priming), if pseudo-alignment is used to generate transcript compatibility counts (Nam et al., 2002; Ntranos et al., 2019).

Despite what the name suggests, some "single cells" will actually be multiple cells, due to two (or more) cells being captured into the same droplet or well, named "doublets". The doublet rate is determined by concentration of cells in the solution / flow rate into the droplets. By Poisson sampling a flow rate can be calculated to minimise doublets but also balance the cost of empty beads. The doublet rate can be determined experimentally by mixing single cells from two different species, in which case double species encapsulations can be detected by the different gene sequences (and the total number of doublets calculated as double that) (Macosko et al., 2015). Doublets can also be detected by natural genetic variation when multiplexing multiple genetically distinct individuals of the same species (H. M. Kang et al., 2018). Artificial genetic variation can also be added by using oligonucleotide-tagged antibodies (Stoeckius et al., 2018). Recently a method using whole-transcriptome pre-indexing was proposed, enabling overloading of droplets and a corresponding massive increase in throughput (Datlinger et al., 2019). A

number of computational tools have also been created to attempt identification of doublets by simulating artificial doublets (Wolock et al., 2019; McGinnis et al., 2019). Another way in which mRNA can cross contaminate across cells is if damaged cells leak mRNA into the solution which is then captured.

1.3.7 Spatial Sequencing

A disadvantage in most single cell RNAseq methods is the loss of spatial information during cell dissociation and isolation. New methods have been developed to combine the spatial advantages of In Situ Hybridisation and the transcriptome wide nature of scRNAseq. One type of method "in situ sequencing" typically uses rolling circle amplification of an mRNA target resulting in a DNA nano-ball which is sequenced by ligation, achieving a resolution of 0.5-1μm for up to 1,000 genes (J. H. Lee et al., 2014; X. Wang et al., 2018; Qian et al., 2018). Another approach "spatial transcriptomics" uses a spatially barcoded bead array which captures mRNA from an overlaid tissue, enabling quantification of all genes but with lower spatial resolution (2-100μm) than in situ sequencing (Ståhl et al., 2016; Rodriques et al., 2019; Vickovic et al., 2019)

1.3.8 Nascent RNA Sequencing

Although we say the previous methods measure "expression" we are actually measuring the total mRNA, which is the cumulative result of transcription minus mRNA degradation. Methods such as GRO/PRO/NET-seq allow profiling of nascent mRNA, revealing which genes are being transcribed at a specific time-point (Core et al., 2008; Churchman et al., 2011; Mahat et al., 2016).

1.3.9 Proteomics

Given that we (mostly) ultimately care about *protein* function, using mRNA as a proxy for this deserves some caveats and justification. Quantification of the full proteome, either by mass-spectrometry approaches or by nanopore protein sequencing, hold promise (with proteins being present in higher copy number than mRNA) but has yet to be fully realised (Reviewed in Specht et al., 2018;

Marx, 2019). Whilst there are many factors resulting in a decoupling of mRNA and protein levels, protein levels are largely determined by mRNA concentrations (Reviewed in Y. Liu, Beyer, et al., 2016).

1.4 Gametogenesis

Gametogenesis is the process by which primordial germ cells develop into either spermatozoa in the male (spermatogenesis) or ovum in the female (oogenesis). Whilst equally important to spermatogenesis, oogenesis is studied far less (1,099 vs 359 articles indexed in PubMed 2018). This is likely partially due to difficulties in studying oogenesis compared to spermatogenesis. In oogenesis prophase I up to diplotene occurs in the ovary of the fetus before birth at which point meiosis is arrested ("Dictyate"), making experiments more challenging.

1.4.1 Testis Anatomy

Whilst the fundamental features of spermatogenesis and oogenesis are the same, spermatogenesis is organised into a specific structure. The testis is composed of seminiferous tubules surrounded by somatic accessory cells such as Leydig cells which produce testosterone. Germ cells are located within the tubule supported by a somatic cell type named Sertoli cells. The germ stem cells are located at the basal surface next to the tubule wall, and more differentiated germ cells are located progressively inwards towards the lumen/apical surface of the tubule (figure 1.3).

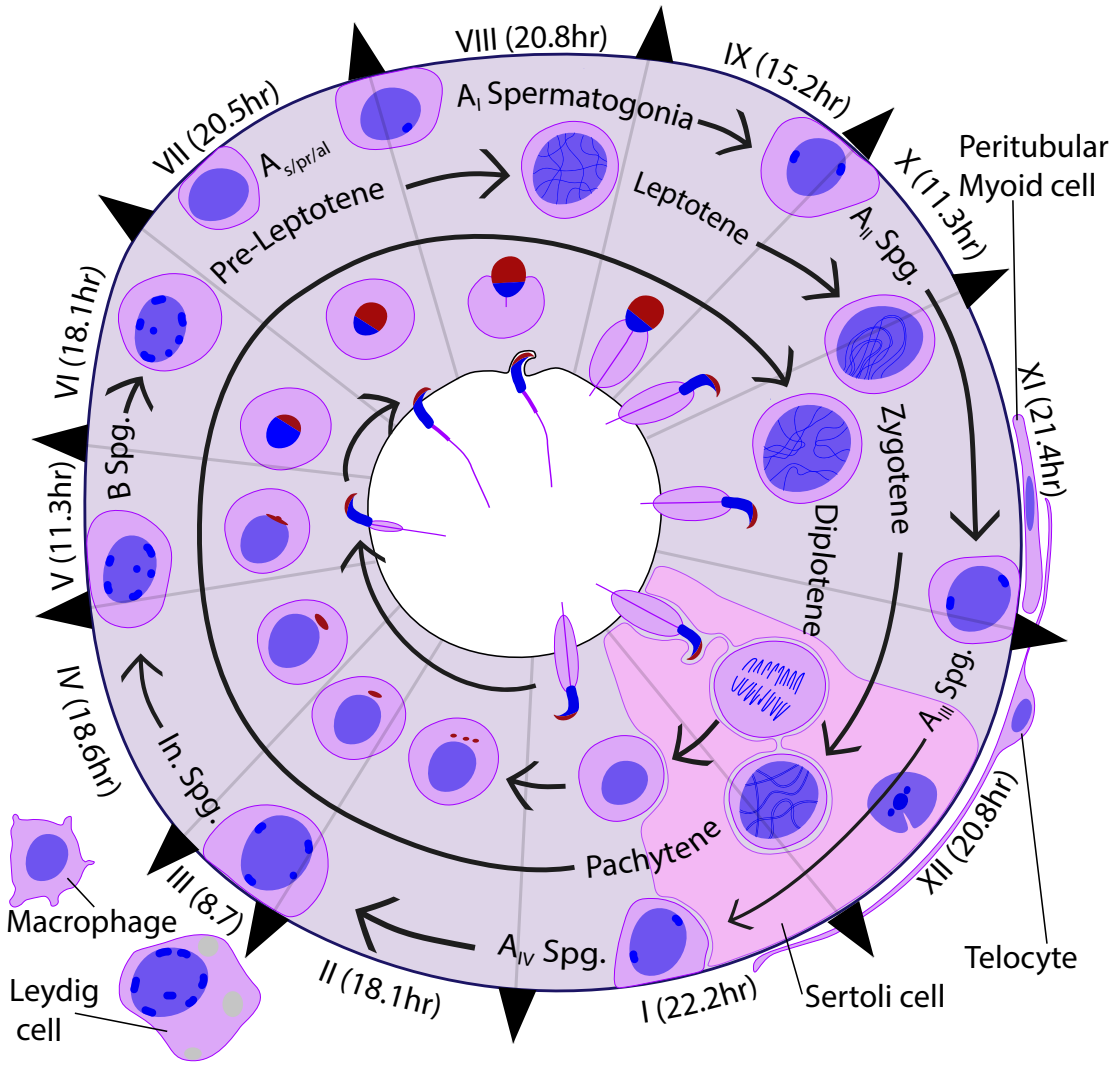


Figure 1.3: The cycle of the seminiferous epithelium shown within a cross section of a seminiferous tubule. To simplify the diagram no intracellular cytoplasmic bridges are shown and only one type of each cell is shown. As/pr/al cells actually exist in all stages. In any given tubule cross section only the cell types from a single stage will be present in mice. Stage numbers are shown around the perimeter of the tubule demarcated by black triangles. In. : Intermediate, Spg. : Spermatogonia. Stage timings from Oakberg, 1956a, Sertoli features from Hess et al., 2015, Chromosome morphology during prophase I from Link et al., 2019, general cycle descriptions from Monesi et al., 1978; Ahmed et al., 2009; Meistrich et al., 2013; Oakberg, 1956b; Hess et al., 2008; Nakata et al., 2015; Muciaccia et al., 2013; Osuru et al., 2014; Lara et al., 2018; Junqueira et al., 2005.

Within a given region of tubule, entry into spermatogenesis is coordinated such that each developing germ cell belongs to a cohort of the same stage. The rate of entry is precisely timed (every 8.6 days in mice) and shorter than the total developmental time (35 days in mice, >70 in humans) such that in any

given tubule section different generations of germ cells occur in layers of particular cell type combinations (figure 1.3, Oakberg, 1956a; Clermont et al., 1969; Heller et al., 1969). This is called the cycle of the seminiferous epithelium and it is possible to stage each combination of cells using periodic acid-fuchsin sulfurous acid staining (14 combinations in rat, and 12 in mouse) (Leblond et al., 1952a; Leblond et al., 1952b; Oakberg, 1956b). The cycle starts at a different time laterally along the tubule (in a so called "wave"), ensuring continuous production of sperm, although the very first wave is synchronous - a property sometimes used to purify or enrich for specific stages.

1.4.2 Spermatocytogenesis

Prior to meiosis there are a number of *mitotic* divisions, amplifying the true spermatogonial stem cell (Oakberg, 1971; Huckins, 1971). The first set of these divisions occur asynchronously between stage stages X and II (Rooij, 2001). There are up to four or five divisions in this first proliferation and the resulting cells are named A_{single}, A_{paired}, $A_{aligned-4}$ $A_{aligned-8}$, $A_{aligned-16}$. At stage VIII all the $A_{aligned}$ cells differentiate (without division) into A_1 cells which are considered committed to the meiotic program and now divide in synchrony with the stages of the seminiferous cycle (Rooij et al., 2000). This commitment step is induced by a pulse of retinoic acid (Vitamin A metabolite), which induces expression of the Stra8 transcription factor in these cells (and pre-leptotene cells which occur in the same stage) (Griswold et al., 2012; Y. Lin et al., 2008; Q. Zhou et al., 2008; Hogarth et al., 2015; Griswold, 2015). A Vitamin A deficient diet results in lack of progression to A_1 spermatogonia, enabling an additional method to generate synchronised cells (J. N. Thompson et al., 1964). A_1 cells divide by mitosis into A_2, then A_3, and A_4 all of which are indistinguishable morphologically but occur at different stages of the seminiferous cycle (Rooij et al., 2000). A_4 cell divide into "In" cells (with intermediate levels of heterochromatin) and then into spermatogonial B cells before a final mitosis generates pre-leptotene spermatocytes (Rooij et al., 2000).

Meiosis is preceded by DNA replication much like in mitosis, and so initiates from cells that are 2n, 4c (n = ploidy / number of centromeres; c = number of chromosomes).

1.4.3 Prophase I of Meiosis I

During meiosis there is an extended prophase I (lasting 14 days in mice), which is itself divided into a number of stages: Leptotene (thin ribbon), Zygotene (union ribbon), Pachytene (thick ribbon), and Diplotene (double ribbon) - originally defined by the morphology of the chromosomes through a light microscope (sometimes suffixed -nema using the greek for "thread" instead of "ribbon") (De Winiwarter, 1900; Grégoire, 1907; Wilson, 1912; Zickler et al., 1998).

However, with the discovery of the synaptonemal complex (figure 1.4, a proteinaceous scaffold that forms along the chromosomes, Moses, 1956; Fawcett, 1956), the stages are also now defined by the configuration of this structure (Zickler et al., 2015). Lepotene being the stage in which the lateral elements form along the axis of homologous chromosomes, Zygotene starts at the formation of the first transverse filaments and the central element between the homologous chromosomes (Moens, 1968), Pachytene is defined as the stage in which all chromosomes (excepting the sex) are fully synapsed, and Diplotene when the synaptonemal complex starts to disassemble (Moses, 1958; Moses, 1977). The lateral elements are formed of SYCP2 (Red1 in *S. cerevisiae*) & SYCP3 proteins, whereas the central element is formed of SYCE1, SYCE2, SYCE3, SIX6OS1, and TEX12, with SYCP1 forming the transverse elements connecting the two (reviewed in Zickler et al., 2015; Gao et al., 2018; Kaniecki et al., 2018; Dunce et al., 2018). The formation of the axis is dependent on a meiosis specific cohesin complex (composed of REC8, RAD21L, SMC1B, STAG3) which encircles the sister chromatids during premeiotic S phase (reviewed in Rankin, 2015; Ishiguro, 2019).

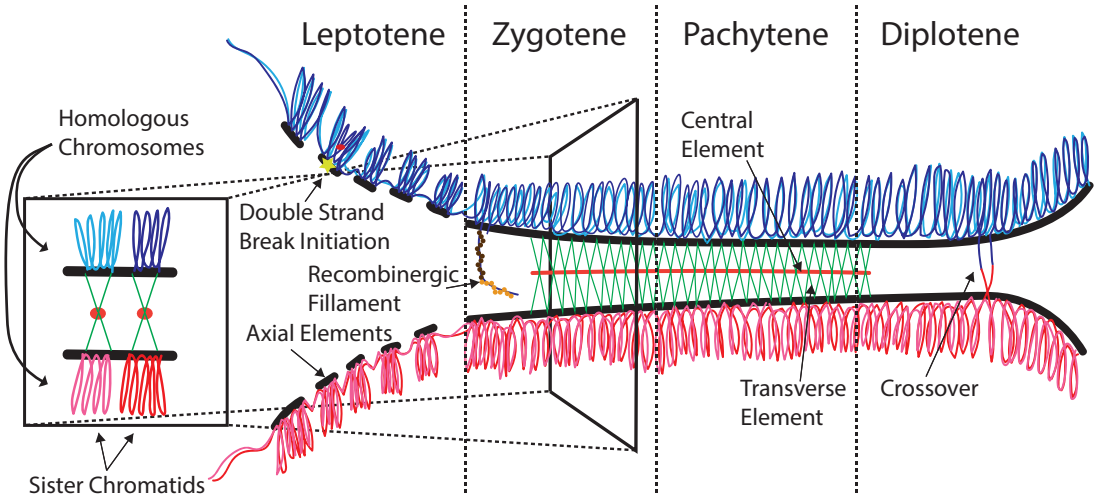


Figure 1.4: Schematic of the synaptonemal complex at different stages during Prophase I (based on Gaysinskaya, B. Miller, et al., 2018; Gaysinskaya, B. F. Miller, et al., 2018; Cohen et al., 2010; Bolcun-Filas et al., 2012; Hughes et al., 2018; Cahoon et al., 2017)

At the start of prophase I in Leptotene, SPO11, a conserved topoisomerase like protein, acts with mTopVIB (GM960 in mice) to create programmed double-strand breaks (DSBs) (H. Sun et al., 1989; Keeney et al., 1997; Bergerat et al., 1997; Cole et al., 2012; Vrielynck et al., 2016; Robert et al., 2016; R. Li et al., 2019). Like other topoisomerases the two SPO11 molecules form covalent bonds with each end of the DNA (Neale et al., 2005).

The activity of SPO11 depends on pre-DSB recombinosomes containing MEI4, REC114, IHO1 (aka CCDC36, MER2) and ANKRD31 which are bound to unsynapsed axes via HORMAD1/2 (Kumar, Bourbon, et al., 2010; Panizza et al., 2011; Stanzione et al., 2016; Kumar et al., 2018; Papanikos et al., 2019; Boekhout et al., 2019). However most DSBs are inflicted on the DNA within one of the many loops that form a linear array, rather than the DNA secured to the central axis from which the loops emanate. Therefore, before DSBs occur, the sites destined to be broken are brought to the axis to form a tethered loop axis complex (figure 1.4) (Blat et al., 2002; Panizza et al., 2011). In *S. cerevisiae* Spp1 provides the link by binding both H3K4me3 and Mer2 (IHO1), however the mouse orthologue CXXC1 does not appear to play the same role, and so it is unclear which protein(s) provide

this link in mice (Acquaviva et al., 2013; Sommermeyer et al., 2013; Tian et al., 2018). MEI1 is also required for DSBs, and WDR61 is hypothesised to be required as it is a homologue of Ski8/Rec14 which is required in *S. cerevisiae* (reviewed in Kumar and De Massy, 2010; I. Lam et al., 2015a). pre-DSB recombinosome recruitment also depends on meiotic cohesin (Bhattacharyya et al., 2019).

The DSB lesions are initially processed in much the same way as DSBs in mitotic cells: they are rapidly recognised by CtIP-(SAE2)-activated MRN complex (MRE11, RAD50, NBS1), which possesses 3'-5' nuclease activity. They are then further resected by either the 5'-3' exonuclease EXO1, or DNA2 nuclease with the BTRR complex (BLM (SGS1) helicase, TOP3A, RMI1, RMI2 (Daley et al., 2014)) resulting in the release of SPO11 bound oligos (Reviewed in Symington, 2014; Schiller et al., 2014; I. Lam et al., 2015a)). MRN recruits the ATM kinase which in turn activates a cascade of DNA damage response activities including phosphorylation of the histone variant H2AX (Maréchal et al., 2013). RPA then coats the single stranded DNA, but is replaced by RAD51 and the meiosis specific protein DMC1, with the aid of BRCA2 (J. Liu et al., 2010; Jensen et al., 2010; Shi et al., 2019). MEIOB and SPATA22 are RPA homologues that are also recruited to DSBs, but their function is unknown (Y. Xu et al., 2017).

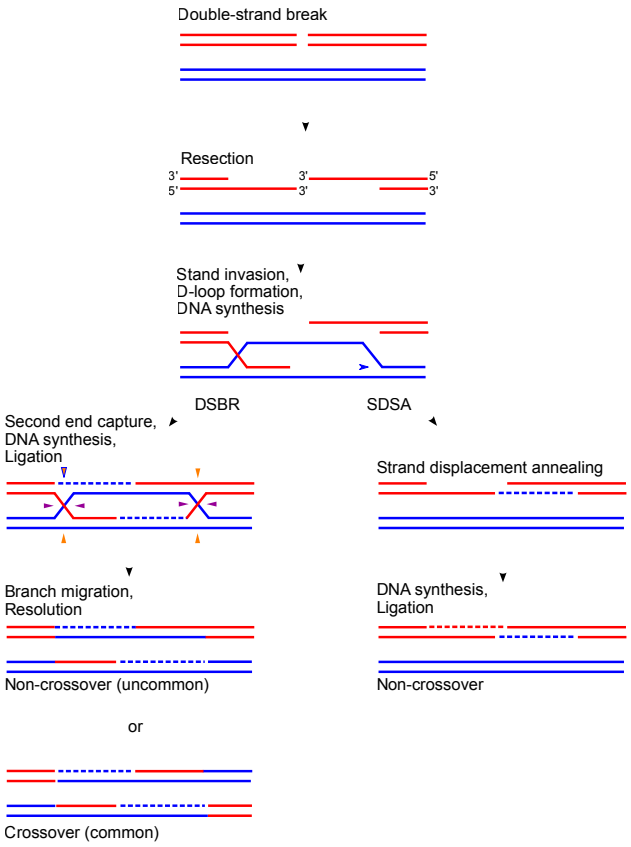


Figure 1.5: Schematic of the recombination process. By Wikipedia user User:Emw2012 (CC BY-SA 3.0), based on Sung et al., 2006

RAD51 and DMC1, along with other helper proteins, form a right handed bipolar helical filament (where DMC1 is oriented at the broken end, with 3bp per monomer), able to invade the homologous chromosome to search for a homologous template for repair (Sehorn et al., 2004; Cloud et al., 2012; Brown, Grubb, et al., 2015; Brown and Bishop, 2015; Crickard et al., 2018). This filament is brought / binds to duplex DNA (perhaps helped by RAD51AP1, PALB2, or HOP2-MND1, Petukhova et al., 2005) and probes for homology using base-flipping - where the nucleoside is rotated out of the double helix (R. C. Gupta et al., 1999; Folta-Stogniew et al., 2004). This search likely involves a combination of 3D search, 1D sliding, and length dependent (>8bp) recognition of micro-homology, reminiscent of Cas9 search (Ragunathan et al., 2012; Forget et al., 2012; Renkawitz et al., 2013; Z. Qi et al., 2015; J. Y. Lee et al., 2015), (reviewed in Barzel et al., 2008; Renkawitz et al., 2014;

Greene, 2016; Kaniecki et al., 2018; Haber, 2018). The filament is likely released from non-homologous interactions by ATP hydrolysis (J. Y. Lee et al., 2016), and is stabilised by MCMDC2 for homologous interactions (McNairn et al., 2017).

At this point the homologous chromosomes are aligned at a distance of approximately 400nm with the invading filament visible as an interaxis bridge. As the synaptonemal complex forms, nucleated at sites designated to become crossovers, the axes are brought closer together to a distance of around 100nm (reviewed in Zickler et al., 2015).

Successful single strand invasion causes a displacement (D) loop (coated by RPA), and a Holliday junction (figure 1.5). DNA synthesis proceeds using the homologous template DNA, possibly by polymerase eta (McIlwraith et al., 2005; Kawamoto et al., 2005) and then branch migration (catalysed by RAD54, RECQ1, and BLM) causes either D loop extension or dissociation (Bugreev et al., 2006; Bugreev et al., 2008; Mazina et al., 2012). In the case of dissociation, the newly synthesised strand can re-anneal to the resected DNA overhang to be repaired by further DNA synthesis and ligation, forming a non-crossover gene conversion (a processed called Synthesis Dependent Strand Annealing, figure 1.5). In the case of loop extension, second end capture (catalysed by RAD52, Sugiyama et al., 1998; Sugiyama et al., 2006; McIlwraith et al., 2008; Lao et al., 2008) followed by DNA synthesis and ligation results in a double Holliday Junction (dHJ). This can also result in a non-crossover gene conversion in a process called double-junction dissolution, in which convergent branch migration is followed by de-catenation (catalysed by BLM-TOPO3A, L. Wu et al., 2003; Bizard et al., 2014).

MUS81-EME1/EME2 can cleave dHJs, as can SLX1-SLX4 and GEN1 when there is no nuclear envelope (reviewed in Wyatt et al., 2014). These DNA nickases can generate both crossovers and non-crossovers depending on whether they nick symmetrically or asymmetrically respectively (figure 1.5). In meiosis the MLH1-MLH3 heterodimer (MutLγ) is recruited to sites with MutSγ, and can resolve the dHJ into crossover products by asymmetric nuclease activity (Zakharyevich

et al., 2012) and (reviewed in Hunter, 2015; Gray et al., 2016; Manhart et al., 2016; Toledo et al., 2019).

MutSγ is a member of a group of proteins, named ZMM proteins, which help to stabilise double Holliday junction intermediates to promote the crossover pathway. This includes MSH4 and MSH5 that as a heterodimer (MutSγ) form a sliding clamp around HJs, HFM1 (MER3), SYCP1 (ZIP1), SHOC1 (ZIP2), SPO16, TEX11 (ZIP4), HEI10 and RNF212 (ZIP3) (reviewed in Pyatnitskaya et al., 2019).

The locations at which these double-strand breaks occur are not uniformly distributed across the genome, although the nonuniformity and locations vary by species. *Caenorhabditis elegans* has quite a uniform distribution with little fine scale structure (Gini coefficient of 0.28) (Rockman et al., 2009; Kaur et al., 2014). *Drosophila melanogaster* is similar with a Gini coefficient of 0.5 (Chan et al., 2012; Smukowski Heil et al., 2015). In other species recombination occurs at a small subset of the genome in clusters called "recombination hotspots", originally named from work on the Chi sequence which can specify recombination in E. coli (S. T. Lam et al., 1974; R. S. Myers et al., 1994). *Saccharomyces cerevisiae* has a Gini coefficient of 0.64, with recombination occurring in hotspots that are 50-300bp, typically (88%) overlapping promoters (in contrast to Drosophila), and are nucleosome depleted (Baudat et al., 1997; T. C. Wu et al., 1994; Mancera et al., 2008; Pan et al., 2011; I. Lam et al., 2015b).

Recombination in humans occurs in hotspots 1-2kb in width that, unlike yeast, are not usually at promoters (Jeffreys et al., 2001). Genome wide locations of these hotspots has been inferred from increasingly high resolution genetic maps derived from linkage disequilibrium patterns (McVean et al., 2004; S. Myers et al., 2005; The International HapMap Consortium, 2007; Coop et al., 2008) and pedigrees which enable sex-specific inference (Kong et al., 2002; Kong et al., 2010; Kong et al., 2014; Bhérer et al., 2017; Halldorsson et al., 2019),

These maps of hotspots lead to the discovery of a 7-mer (CCTCCCT) associated with recombination rate (S. Myers et al., 2005), which appeared to be causal (Jeffreys et al., 2002). This motif was later extended to the 13-mer CCNCCNTNNCCNC

(S. Myers et al., 2008) which was found to match the predicted binding motif of PRDM9 (S. Myers et al., 2010; Baudat et al., 2010; Parvanov et al., 2010; Berg et al., 2010). In addition to the zinc finger array (which specifies the binding motif of the protein), *Prdm9* also has a KRAB domain, an SSXRD domain, and a PR/SET methyltransferase domain. The methyl-transferase domain is able to catalyse both H3K4me3 and H3K36me3, and both of these marks are required for its role in positioning DSBs (Powers et al., 2016; Diagouraga et al., 2018).

Prdm9 is the founding member of a family of KRAB zinc finger genes (Birtle et al., 2006). There are around 600 such genes in mice (only ~8 in birds), most of them likely to be involved in repressing retrotransposons by recruiting KAP1 (aka TRIM28) but PRDM9 does not have this ability (Imbeault et al., 2017; Bruno et al., 2019). *Prdm9* is among the fastest evolving regions of the genome (Oliver et al., 2009), and consistent with this the locations of hotspots in Chimp are not shared with Humans (Wall et al., 2003; Ptak et al., 2004; Ptak et al., 2005; Winckler et al., 2005). However, general features of the recombination landscape are the same in Chimps, including increasing heat at the sub-telomeres, hotspot width and intensity, and rate depression within gene bodies (Auton et al., 2012). Similar results have also been obtained in mice (Paigen et al., 2008; Brunschwig et al., 2012; Booker et al., 2017), with some studies utilising the increased feasibility of experimental manipulation of mice to use alternative detection methods such as assaying the intermediates of recombination such as ssDNA bound DMC1 (Smagulova et al., 2011; Khil et al., 2012; Smagulova et al., 2016), SPO11 oligos (Lange et al., 2016), as well as using single sperm sequencing of hybrids with genetic differences (much greater than in humans)(Hinch et al., 2019), and linked reads (Dréau et al., 2019). DMC1 mapping has also been applied to human samples revealing the contribution of individual *Prdm9* alleles (Pratto et al., 2014a) and recently low coverage high throughput single sperm sequencing has also been used on human samples (Bell et al., 2019).

In Canids recombination rate is even more unequally distributed than in humans, is similarly increased towards the subtelomeres (especially in males), but is directed towards CpG rich sequences such as promoters as in yeast (Auton et al., 2013;

C. L. Campbell et al., 2016); consistent with the loss of functional *Prdm9* in the Canidae family (Muñoz-Fuentes et al., 2011). Similarly in birds, which also lack a functional *Prdm9*, recombination hotspots exist but are conserved across species diverged by tens of millions of years and are enriched at CpG islands including transcription start sites - perhaps driven by nucleosome depletion and so increased accessibility to the recombination machinery (Singhal et al., 2015). Plants also lack *Prdm9*, and again crossovers are enriched at promoters, and other promoter associated marks such as H2A.Z and H3K4me3 (K. Choi et al., 2013; Marand et al., 2017).

Furthermore, knockout of *Prdm9* in mice causes hotspots to relocate to promoters (Brick et al., 2012). *Prdm9* B6 knockout mice are infertile, but this is not true for other strains, and one *Prdm9* homozygous null female human has had three healthy children (Hayashi et al., 2005; Mihola et al., 2019; Narasimhan et al., 2016). A partial relocation to promoters is seen in *Ankrd31* knockouts (Boekhout et al., 2019; Papanikos et al., 2018). *Prdm9* is also the cause of infertility in the F1 male hybrids of crosses between *Mus musculus domesticus* (B57BL/6J) and *Mus musculus musculus* (PWD/Ph, differing by only a single finger), representing the only known mammalian speciation gene (Forejt et al., 1974; Mihola et al., 2009). The cause of this sterility was recently revealed by engineering mice with a humanized version of *Prdm9*, which rescues the fertility of F1 hybrids (B. Davies et al., 2016). Hotspots are gradually eroded by biased gene conversion (the colder allele is used as a template to repair the double strand break occurring on the hotter template), resulting in the paradox of hotspot existence (Boulton et al., 1997; Coop et al., 2007; S. Myers et al., 2010; C. L. Baker et al., 2015). Using a humanised zinc finger increases the symmetry of DSB formation, and this increased symmetry is associated with increased synapsis and fertility. Additionally asymmetric hotspots have greater DMC1-ChIPseq signal compared to H4K3me3 (a measure of PRDM9 activity) suggesting delayed repair at PRDM9-asymmetric sites. This work also implies that PRDM9 plays a role not just in DSB positioning but DSB repair in a manner depending on symmetric binding. Further work using hybrid mice

pedigrees and single sperm sequencing demonstrated that at sites with asymmetric PRDM9 binding, homologue templated repair was depleted compared to PRDM9-symmetric sites, showing that PRDM9 influences repair as well as positioning of DSBs (R. Li et al., 2019; Hinch et al., 2019).

While there are many potential binding sites for PRDM9 (>170,000 in a HEK293T overexpression system (Altemose et al., 2017)) and hence potential crossover locations, there are only tens of thousands of hotspots (S. Myers et al., 2005; Pratto et al., 2014a), and only 1-2 crossovers form per chromosome in a single meiosis. At each stage of recombination the number of potential crossover sites is reduced, and at each stage various factors influence the likelihood of a site surviving to the next stage of selection. While PRDM9 likely deposits H3K4me3 at almost all locations it binds, at some regions such as promoters recombination is suppressed, plausibly due to poor H3K36me3 deposition or active removal of this mark (Altemose et al., 2017). Sequence motifs outside the PRDM9 binding footprint also influence recombination, some acting to repress recombination by recruitment of KRAB-ZNF proteins which recruit TRIM28 which in turn deposits H3K9me3 (Altemose et al., 2017). H3K9me3 is a marker of heterochromatin, which is more often found in the gene poor 'B' compartments of chromosomes and laminin associated domains, and is associated with lower PRDM9 binding and recombination (Walker et al., 2015; Patel et al., 2019; Yamada et al., 2017). Furthermore few if any DSBs occur in the pericentromeric heterochromatin (Verver et al., 2013), and disruption of H3K9me2 and non-CG DNA methylation pathways in *Arabidopsis thaliana* results in increased recombination at the previously heterochromatic centromeres (Underwood et al., 2018). In mice SPO11 creates ~250/300 DSBs in males/females (Baudat et al., 2007). However these numbers and direction of sex bias vary by species, for example in Lily over 2,000 DSBs are formed (as assessed by RAD51 foci counts) (Terasawa et al., 1995), while in species able to pair homologues without recombination (*Drosophila melanogaster* and *Caenorhabditis elegans*) only 20-30 breaks are formed. Resolution of a DSB as a crossover is more likely if it occurs closer to the telomere, has higher local GC content, has higher H3K4me3, and has more symmetric binding of PRDM9

(Hinch et al., 2019). Additionally, crossovers are prevented from occurring too close to one another (interference), and at least one crossover is required per chromosome (assurance/obligate) (Fledel-Alon et al., 2009; Hunter, 2015; Otto et al., 2019).

1.4.4 Post-Pachytene Processes

Once repair has completed without the pachytene checkpoint being activated, the synaptonemal complex dissociates and cells undergo two divisions without an intervening S phase, the first reductional (separating homologous chromosomes), the second equatorial (separating sister chromatids) (Roeder et al., 2000; Subramanian et al., 2014; Watanabe, 2012).

Spermatocytes then begin the biochemical and morphological changes to develop into a spermatid. One major change is the formation of the acrosome (reviewed in Buffone, 2016; Khawar et al., 2019). In the first phase (Golgi phase, steps one to three), the Golgi creates proacrosomal vesicles that fuse to become the acrosomal vesicle (although some are created as early as late pachytene). At step three the anexome (a cytoskeletal structure responsible for flagella beating) begins to form from the distal centriole (figure 1.6). There is both a proximal and distal centriole below the nucleus, each of which is a microtubule based structure of 9 radial triplet microtubules arranged in a barrel (Fawcett et al., 1969; Avidor-Reiss et al., 2019). The anexome consists of a central pair of microtubules encircled by nine doublet microtubules connected by Nexin ("9+2" configuration, figure 1.6) (Linck et al., 2016; Lehti et al., 2017). Dynein motors slide these microtubules against each other creating the beating motion required for sperm motility (Mitchison et al., 2010).

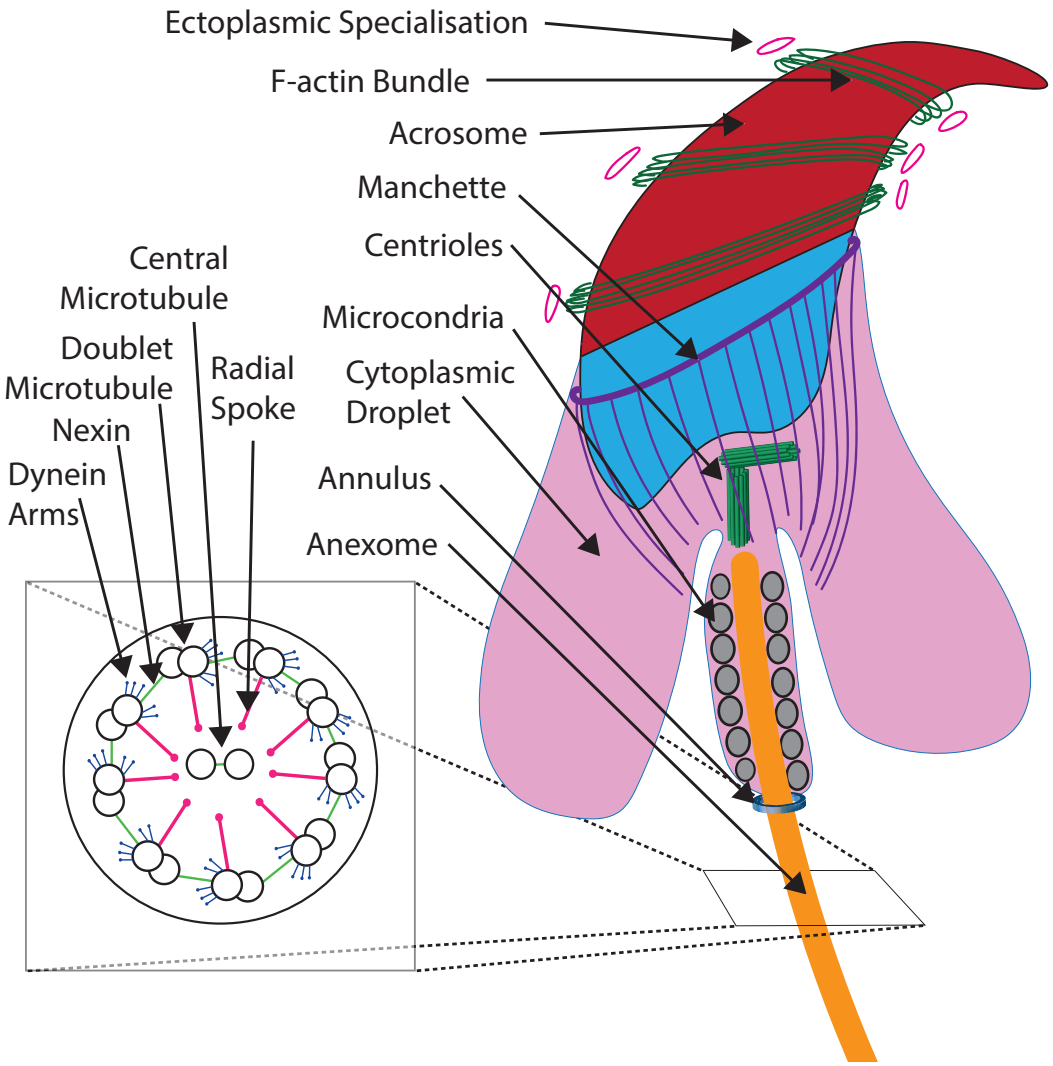


Figure 1.6: Schematic of subcellular structures in spermatids. Based on (Dunleavy et al., 2019; Y.-L. Wei et al., 2018; O'Donnell et al., 2012; Kopera et al., 2010; W.-L. Wang et al., 2019; Gunes et al., 2020)

In the cap phase (steps four to seven) the acrosomal vesicle flattens to cover one side of the nucleus where it becomes anchored to the nuclear envelope by an F-actin and SAK57 (keratin5 orthologue) containing cytoskeletal structure called the acroplaxome (Kierszenbaum et al., 2003; Kierszenbaum et al., 2004).

In the acrosome phase (steps eight to twelve) the acrosome orients towards the base of the tubule in close apposition to the plasma membrane. In addition the apical ectoplasmic junction forms, anchoring the sperm head to the Sertoli cell (E. W. P. Wong et al., 2008). The transient skirt like structure of the manchette

also forms, consisting of a perinuclear ring important for sperm head shaping, and microtubules emanating towards the tail which are important for protein transport (figure 1.6, Lehti et al., 2016). Mitochondria are arranged in a helical sheath in the midpiece of the sperm, terminated at the base by the annulus which is composed of Septin octamer fibres associated with TAT1 and through which the axoneme passes (Ho et al., 2007; Toure et al., 2011; Kuo et al., 2015). Nuclear condensation begins with up to 99% (90%) of mice (human) histones being replaced with transition proteins and then protamines. These short (50 a.a.) highly basic (~50% arginine) proteins form a more compact toroid structure enabling a 90% reduction in nuclear volume (Balhorn, 2007; Ward, 2010; Yamaguchi et al., 2018). This nuclear condensation results in transcriptional cessation and therefore requires many mRNAs to be stored for long periods of time. For example *Prm1* (Protamine 1) and *Smcp* (Sperm Mitochondria-associated Cysteine-rich Protein) mRNA are stored for three and seven days respectively as free ribonucleic particles before they are translated into protein on polysomes (Cullinane et al., 2015; Kleene et al., 1984; Kleene, 2004).

In the maturation phase nuclear condensation continues while excess cytoplasm is shed as residual bodies which are digested by Sertoli cells (figure 1.6, Lacy, 1962; Breucker et al., 1985; Hermo et al., 2010). Spermatids are then released where they mature fully within the epididymis. Before fertilisation the acrosome reaction occurs in which the outer acrosomal membrane fuses with the sperm plasma membrane, releasing the hydrolytic acrosomal enzymes to aid in fertilisation (M. Jin et al., 2011).

1.5 Prior Work

A number of previous studies have measured spermatogenic gene expression. Some studies such as GTEx (Genotype-Tissue Expression), HPA (Human Proteome Atlas), and FANTOM (Functional Annotation Of Mammalian genome) consortia have transcriptionally profiled many whole tissues in humans using RNAseq (Mele et al., 2015; Uhlén et al., 2015; Uhlén et al., 2016). This data enables the identification of tissue specific genes and hence a preliminary list of proteins which could be

involved in meiosis. However, the testis is a complex amalgamation of different cell types and bulk tissue sequencing can generally not resolve the gene expression for different cell types instead resulting in a proportion weighted average. The relative inability to faithfully culture mammalian meiotic cells in vitro has also hampered the ability to generate higher resolution time-course data (Q. Zhou et al., 2016; Hikabe et al., 2016).

Nonetheless a number of techniques have been developed to improve on whole tissue resolution. For example the initial wave of spermatogenesis is somewhat synchronous and so sampling during this time could in theory yield homogeneous cell types/stages. However it is likely there is some variation in synchronicity resulting in impure samples and uncertain classification (Laiho et al., 2013; Ball et al., 2016). Some studies have combined this approach with an in silico deconvolution or machine learning algorithms (Margolin et al., 2014; Y. Li et al., 2013). One study utilised the greater synchrony of meiosis in fetal ovaries in combination with germ cell depleted mutant mice to partially delineate the meiotic prophase gene regulatory programme (Soh et al., 2015). Other approaches include either size based centrifugal sorting (Soumillon et al., 2013; Buard et al., 2009; Grabske et al., 1975) or FACS (Fluorescence Activated Cell Sorting) using a DNA stain (da Cruz et al., 2016), both of which result in a limited number of (possibly quite heterogeneous) cell populations. Immunohistochemistry enables single cell resolution of protein expression but is low throughput (Djureinovic et al., 2014).

Knowledge of the yeast gene regulatory programme of meiosis is relatively well known, however, the lack of sequence similarity and inability to culture mammalian meiotic cells has hampered efforts to translate this network to more complex eukaryotes (Brar et al., 2011; Mata et al., 2002; Chu et al., 1998; Handel et al., 2010).

1.6 Statistical Methods

Although the methods and workflow for analysing bulk RNAseq are now fairly mature and standardised, scRNAseq data has characteristics such as sparsity and high number of cells which are poorly handled by bulk RNAseq tools. In addition

scRNAseq data has opened up new analysis possibilities such as pseudotemporal ordering which is not possible with bulk data. This has motivated the development of many (538 in a curated database as of 18th December 2019, Zappia et al., 2018) new methods and analysis packages to meet the new demands and challenges of analysing scRNAseq data (Zappia et al., 2018). In fact, the total number of tools to analyse single cell RNAseq data is comparable to the total number of single cell RNAseq datasets in a curated database (683) (Svensson, Beltrame, et al., 2019).

There are two major analysis modes for scRNAseq, each of which can be applied to either cells or genes. These two analysis "modes" are actually extremes of an analysis spectrum ranging from continuous to discrete depending on whether cells are considered to be separate groups or as part of a (developmental) trajectory. For the discrete case cells can be clustered into types, and differentially expressed marker genes can be inferred. Examples of these kinds of methods include k-means, hierarchical clustering, density clustering, or consensus clustering (Žurauskienė et al., 2016; Kiselev et al., 2017; M. Guo et al., 2015; Satija et al., 2015). For the continuous case cells can be ordered in a pseudo-timeline (or more generally arranged in lineage trees), and gene expression dynamics inferred. Pseudo-temporal ordering of cells is typically achieved by creating a minimum spanning tree and then projecting cells onto the shortest path to create a timeline (Trapnell et al., 2014; Ji et al., 2016). Alternatively a principal curve / graph or seriation can be used (Marco et al., 2014; Qiu et al., 2017).

1.6.1 Dimensionality Reduction

These analysis methods, however, typically struggle with the extremely high dimensionality of scRNAseq (with each gene representing a dimension), sometimes referred to as the curse of dimensionality (Donoho, 2000). Dimensionality reduction is therefore often performed on the gene by cell count matrix prior to application of these methods. The rationale behind this is that the data does not uniformly fill the high dimensional space, but actually lies on a manifold of lower dimension (the manifold hypothesis, Moon et al., 2018), and so by projecting the cells onto this

manifold the dimensionality can be greatly reduced while preserving much of the true structure in the data. This quasi-high dimensional data arises when there is a simpler underlying data generating process, for example a transcription factor may activate many different genes in the same cells, resulting in highly correlated gene expression and therefore in a sense "redundant" dimensions that provide little additional information. Equivalently, cells exist in a limited number of transcriptional states compared to all the possible states given all possible combinations of gene expression. Dimensionality reduction also enables visualisation of the data in 2 or 3D, greatly enhancing an analysts' intuition when reasoning about a dataset.

For visualisation purposes non-linear dimensionality reduction techniques are most commonly used, especially t-SNE (t-distributed Stochastic Network Embedding), and more recently its derivative UMAP (McInnes, Healy, Saul, et al., 2018; Becht et al., 2019). But other methods are also used such as diffusion map based approaches (PHATE and URD Haghverdi et al., 2015; Moon et al., 2019; Farrell et al., 2018), force directed graph methods like SPRING (Weinreb et al., 2018; Wagner et al., 2018), locally linear embedding (J. D. Welch et al., 2016), and Gaussian process latent variable models (K. Campbell et al., 2015).

However, the lack of interpretable parameters in these methods limits their use for inferring co-expressed gene sets. Linear factorisations are therefore often used (in addition to the non-linear methods), for example Principal Components Analysis (PCA) (Alter et al., 2000; Green et al., 2018), Independent Component Analysis (ICA) (Saunders et al., 2018), Non-Negative Matrix Factorisation (NMF) Brunet et al., 2004; P. M. Kim et al., 2003; C. Shao et al., 2017; X. Zhu et al., 2017; Duren et al., 2018; Kotliar et al., 2018; J. D. Welch et al., 2019 and Factor Analysis (Bernardo, 2003; Buettner et al., 2015, reviewed in Stein-O'Brien et al., 2018).

NNMF is often motivated by the positive nature of the original data, in addition to potentially increased interpretability for purely positively additive components which also naturally have a degree of sparsity in both the cell scores and gene loadings due to the positivity constraint. However, we note that latent factors which allow negative loadings have the potential to better capture transcriptional repression.

These methods infer linear subspaces / hyperplanes in the high dimensional data enabling projection of the data and dimensionality reduction. Each component / factor represents a mode of variation in the data, which could arise due to either a biological or a technical process. The data is modelled as a weighted mixture of these components represented by the product of two lower rank matrices, just as the measured data represent the contribution of multiple underlying (latent) processes. The row vectors of the first matrix specify for each cell the mixing weights of each factor. The column vectors of the second matrix specify the contribution of each gene to the components. Put another way the column and row vectors of each matrix specify how active each cell and gene is for a given component. While these methods are typically less useful for visualisation (as the more restrictive linear structure limits the amount of structure that can be represented by 2 or 3 factors) the cell scores and gene loadings provide a direct way to interpret the low dimensional structure with the cell scores effectively soft clustering the cells and the gene loadings giving a set of co-expressed marker genes for each soft-cluster. In addition this linear reduction can still be used as input to the non-linear methods as a second visualisation step.

Some of these methods have been adapted specifically to account for the sparsity of scRNAseq datasets, the idea being that some technical effects caused "dropouts" that should be corrected or accounted for. ZIFA (Zero Inflated Factor Analysis) is one example, which accounts for the sparsity by including a zero-inflation component in the model (Pierson et al., 2015). However, it is likely this apparent inflation of zeros is actually an artefact of normalising count data and a negative binomial model fits well for tag based scRNAseq datasets (Vieth et al., 2017; Svensson, 2019; Townes et al., 2019).

Beyond matrix factorization, there are other frameworks with similar goals that have been applied successfully to single cell data. One set of methods are those based on neural networks, such as self-organizing maps (Löffler-Wirth et al., 2015; D. H. Kim et al., 2015) and auto-encoder neural networks (DCA & scVI) (Eraslan et al., 2019; Lopez et al., 2018). These methods (much like t-SNE) create a

non-linear embedding of the high dimensional data, resulting in a lower dimensional set of scores for each cell. This approach does not, however, provide the equivalent to gene loadings and so an additional differential expression analysis on a hard clustering of the latent embeddings has to be performed in order to find genes associated with the latent dimensions. However, more recently a hybrid approach has been proposed sacrificing some of the flexibility to gain in interpretability (Svensson and Pachter, 2019).

1.6.2 Sparse Decomposition of Arrays

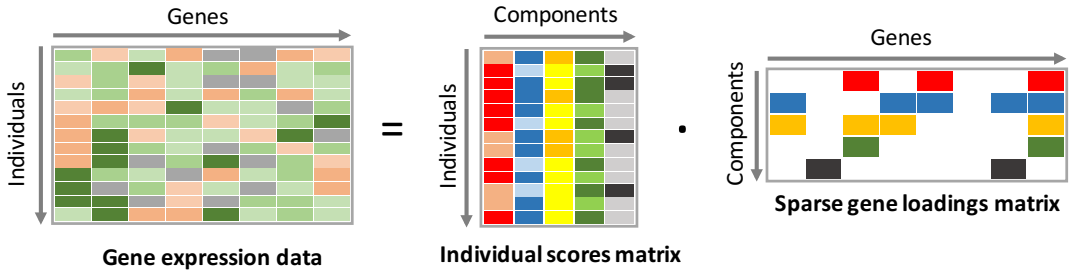


Figure 1.7: Schematic of the SDA matrix factorisation. From (Hore et al., 2015)

Sparse decomposition of Arrays (SDA) is a sparse Bayesian factor analysis method, originally developed for decomposing a tensor of multi-tissue bulk RNAseq data, but can also be used in a group factor decomposition mode or to decompose a single 2D matrix (figure 1.7) (Hore et al., 2015; Hore et al., 2016). The full model for the single matrix decomposition is given in equation 1.1 and in figure 1.8.

$$
\begin{aligned}
y_{nl} &= \sum_{c=1}^{C} a_{nc} x_{cl} + \epsilon_{nl} \\
P(\mathcal{Y}|\theta) &= \prod_{l} \mathcal{N}_N \left(\mathbf{y}_l | \sum_{c} \mathbf{a}_{\cdot c} x_{cl}, \lambda_l^{-1} I_N \right) \\
P\left(a_{nc}\right) &= \mathcal{N}\left(a_{nc}|0,1\right) \\
x_{cl} &= w_{cl} s_{cl} \\
P\left(w_{cl}|\beta_c\right) &= \mathcal{N}\left(w_{cl}|0,\beta_c^{-1}\right) \\
P\left(\beta_c\right) &= \mathcal{G}\left(\beta_c|e,f\right) \\
P\left(s_{cl}|\psi_{cl},\phi_{cl}\right) &= \mathcal{B}\text{ernoulli}\left(s_{cl}|p_{cl}\right) \\
p_{cl} &= \psi_{cl}\phi_{cl} \\
P\left(\psi_{cl}\right) &= \mathcal{B}\,\text{eta}\left(\psi_{cl}|g,h\right) \\
P\left(\phi_{cl}|\rho_c\right) &= \mathcal{B}\text{ernoulli}\left(\phi_{cl}|\rho_c\right) \\
P\left(\rho_c\right) &= \mathcal{B}\,\text{eta}\left(\rho_c|r,z\right) \\
P\left(\lambda_l\right) &= \mathcal{G}\left(\lambda_l|u,v\right)
\end{aligned}
\tag{1.1}
$$

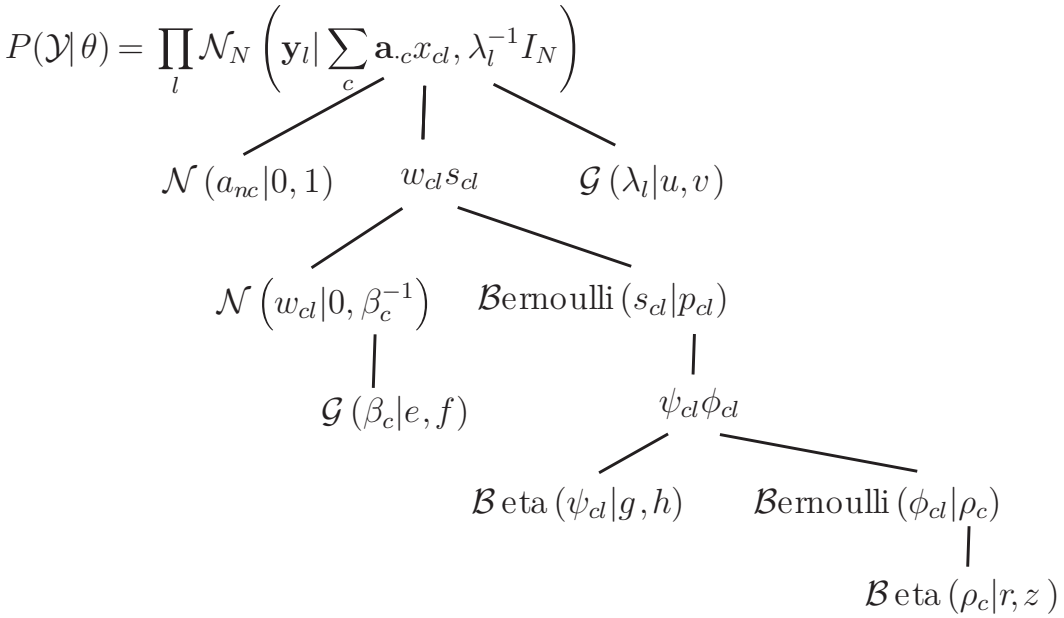


Figure 1.8: Schematic of the SDA Hierarchical Bayesian Model

$\mathcal{Y}$ is an N by L matrix of cells by genes, $\mathcal{A}$ is a matrix of N cells by C

components, and $\mathcal{X}$ is a matrix of C components by L genes, and $\mathcal{E}$ is an N by L matrix of independent Gaussian noise.

A key feature of this model is the element wise sparsity of the gene loadings which eases interpretation of the model, forms a kind of regularisation, and prevents the solution from being rotationally invariant. Moreover this reflects a prior that generally only a subset of genes should be involved in any one biological process. This sparsity is achieved by using a spike and slab prior which is a mixture of a point mass at 0 (spike) and a normal distribution (slab). The slab w has a precision of β_c which has a Gamma prior with hyper-parameters $e = 10^{-6}$ and $f = 10^{-6}$ (diffuse). The spike s is modelled as a Bernoulli distribution with probability parameter p_{cl} representing the probability of an individual gene loading being non-zero. The prior on this parameter is itself a spike and slab with the slab ψ_{cl} having a Beta prior with hyper-parameters $g = 0$ and $h = 0$ (a prior with half mass at 0 and half at 1). The spike ϕ_{cl} again has a Bernoulli prior with parameter ρ_c which also has a Beta prior with hyper-parameters $r = 1$ and $z = 1$ (flat). This hierarchical structure encourages sparsity as when ρ_c is close to 0 the prior on the main spike and slab will be dominated by the spike. The factored spike and slab is equivalent to the pure mixture as in equation 1.2 but aids inference.

$$P\left(x_{cl}|p_{cl},\beta_c\right) = p_{cl}\mathcal{N}\left(x_{cl}|0,\beta_c^{-1}\right) + \left(1-p_{cl}\right)\delta_0\left(x_{cl}\right) \tag{1.2}$$

Each cell score has a standard normal prior. Due to scaling indeterminacy (for example the cell scores could be multiplied by two and gene loadings halved with equivalent fit), the variances are set to 1 with any scaling occurring in the β_c of the gene loadings. The sign of the components is also unidentifiable, in addition to permutation of the components.

Each gene has a noise precision parameter λ_l with a Gamma prior where the hyper-parameters are set to $u = 10^{-6}$ and $v = 10^{-6}$ resulting in a diffuse prior.

The model is fit using variational Bayes as described in (Hore et al., 2015; Hore et al., 2016). As variational Bayes underestimates variances, the mean of the posterior distributions are used rather than the distributions themselves (Jaakkola

et al., 2000). Sparse factor analyses using the spike and slab prior have also previously been used (West, 2003; Lucas et al., 2006). The complexity is $\mathcal{O}(NLC^2)$ and the number of components must be chosen ahead of time, although in practice SDA can set entire components to zero if too many components are used (Hore et al., 2015; Hore et al., 2016).

Wir haben gesehn, dass alle Organismen aus wesentlich gleichen Theilen, nämlich aus Zellen zusammengesetzt sind.

We have seen that all organisms are composed of essentially similar parts, namely, of cells.

— Schwann, 1839

2

Single Cell Transcriptomics

Contents

2.1 Introduction 38
2.2 Experimental Data Generation 39
2.3 Data Processing and QC 39
2.4 SDA 40
2.5 Low Dimensional Visualisation 45
2.6 Imputation 51
2.7 Mutant Genotypes 54
2.8 Potential Artefactual Components 59
2.9 Components 60
2.9.1 Telocytes 62
2.9.2 Leydig Cells 62
2.9.3 Sertoli Cells 63
2.9.4 Macrophages 63
2.9.5 Lymphocytes 64
2.9.6 Peritubular Myoid 64
2.9.7 Spermatogonia 64
2.9.8 (pre)Leptotene & Zygotene 65
2.9.9 Pachytene 67
2.9.10 Hormad1 Sex Chromosome Activation 71
2.9.11 Cul4a, Mlh3, and Cnp Mutants 73
2.9.12 Meiotic Divisions 75
2.9.13 Round Spermatid 76
2.9.14 Spermiogenesis 81
2.10 Motif Inference 83
2.10.1 Enrichment of Known Transcription Factor Targets . . . 83
2.10.2 *De-novo* inference from promoter sequences 84
2.10.3 Broad-scale patterns of TFBS occurrence 85
2.10.4 Novel Motifs 88

2.11 Interactive Web App 88
2.12 Detailed Methods 90
2.12.1 Preprocessing of Drop-seq data 90
2.12.2 Software 90
2.12.3 X Chromosome Non Sharing Simulation 91
2.12.4 Imputation Validation 91
2.12.5 Transcription Factor Motif Analysis 92
2.12.6 Data availability 94

2.1 Introduction

Cells are the fundamental unit of all living organisms, and with recently developed technologies such as DropSeq it is now possible to perform mRNA sequencing of individual cells for tens of thousands of cells. This enables discovery and characterisation of novel cell types, clues towards gene function via co-expression, as well as characterisation of transcriptional processes such as transcription factor binding.

Here we have generated a dataset of >20,000 cells from both wild type and knockout mice testis, using a matrix factorisation approach to soft cluster cells, infer corresponding co-expressed gene markers, and impute noisy gene expression. This revealed novel marker genes, novel predicted functions for understudied genes, further characterisation of understudied cell types, pathology associated process, and insights into sex chromosome inactivation. We also inferred, *de novo* transcription factor binding motifs, revealing a haploid switch in the types of motifs used.

All experimental work in this chapter was performed by Min Jung and colleagues in Don Conrad's group at the University of Washington, St Louis. All analysis steps were performed by myself unless otherwise stated. Much of this chapter has been published as Jung et al., 2019, on which I am co-first author.

2.2 Experimental Data Generation

The cells were dissociated by two methods: enzymatic for the first two wild type mice and the spermatogonia, and mechanical for all other samples (Lima et al., 2017; Jung et al., 2019). The cells were then encapsulated using the DropSeq protocol (Macosko et al., 2015) as described in Jung et al., 2019.

The samples generated are as follows:

- 11 wild type C57BL/6 mice samples
- 3 Hoechst-FACS samples from a 12th wild type (primary spermatocyte (4C), secondary spermatocyte (2C), and spermatid(1C))
- 1 sample of FACS sorted spermatogonial cells from 5 wild type C57BL/6 litter mates with GFP tagged Pou5f1 (B6;CBA-Tg(Pou5f1-EGFP)$^{2Mnn/J}$, from Szabó et al., 2002)

In addition to wild type mice four different mutant mice were included:

- 6 *Mlh3*$^{-/-}$ mice (B6.129-Mlh3^{tm1Lpkn}/J, Lipkin et al., 2002)
- 2 *Hormad1*$^{-/-}$ mice (B6;129S7-Hormad1^{tm1Rajk}/Mmjax, Shin et al., 2010)
- 2 *Cul4a*$^{-/-}$ mice (B6;129-Cul4a$^{-/-}$, Yin et al., 2011)
- 2 CNP eGFP BAC TRAP (knockin) C57BL/6 mice (Joseph Dougherty).

2.3 Data Processing and QC

The samples were sequenced on either Illumina HiSeq2500 or MiSeq. Reads were mapped to the mouse genome using DropSeq tools provided by Macosko lab (using STAR aligner and GRCm38 release 90).

The resulting cell-gene count matrices were merged (54,251 cells and 38,317 genes) and processed through a series of quality control and normalization steps:

Genes with a unique molecular identifier (UMI) count of less than 5 or being expressed in fewer than 5 cells were removed.

Cells meeting the following criteria were removed:

- UMI count of less than 200
- Fewer than 100 genes expressed
- Log UMI count more than 1 standard deviation below the mean for that experiment
- Log number of genes expressed was more than 1 standard deviation below the mean for that experiment

A t-SNE dimensionality reduction of the filtered data revealed an amorphous homogeneous group of cells with low library size, high mitochondrial gene expression and often co-expressing genes from both early and late spermatogenesis suggesting poor quality and/or doublet cells. This group of cells was therefore removed before further analysis, in addition to any cells with a normalised *mt-Rnr-2* expression of greater than 2 suggestive of lysed cells (Ilicic et al., 2016). These filters resulted in 20,322 cells and 28,893 genes remaining.

Genes in the lower third of expression means were then removed and cells were normalized by square root transformation of total transcript counts per cell and genes were normalized to unit variance. All expression values were capped to a maximum of 10. This results in a final matrix of 20,322 cells by 19,262 genes with a sparsity (frequency of 0s) of 93.8% and a median UMI count of 1,312 per cell.

2.4 SDA

When performing a dimensionality reduction such as matrix factorisation the number of latent components must be determined. Too low and real components may be missed, too high and spurious components may be inferred due to overfitting. One way the method can overfit is by assigning a whole component to an individual cell, which we observed (figure 2.1). When increasing the number of components from 50 to, 75, 100, 200 and 500, predominantly these single cell components were inferred, rather than new normal components. With 50 components we inferred

five individual cell components (1,4,8,14, and 46 - figure 2.2) which we judged to be an acceptable balance. We also ran SDA with five different random seeds to confirm stability of the results with different initialisations.

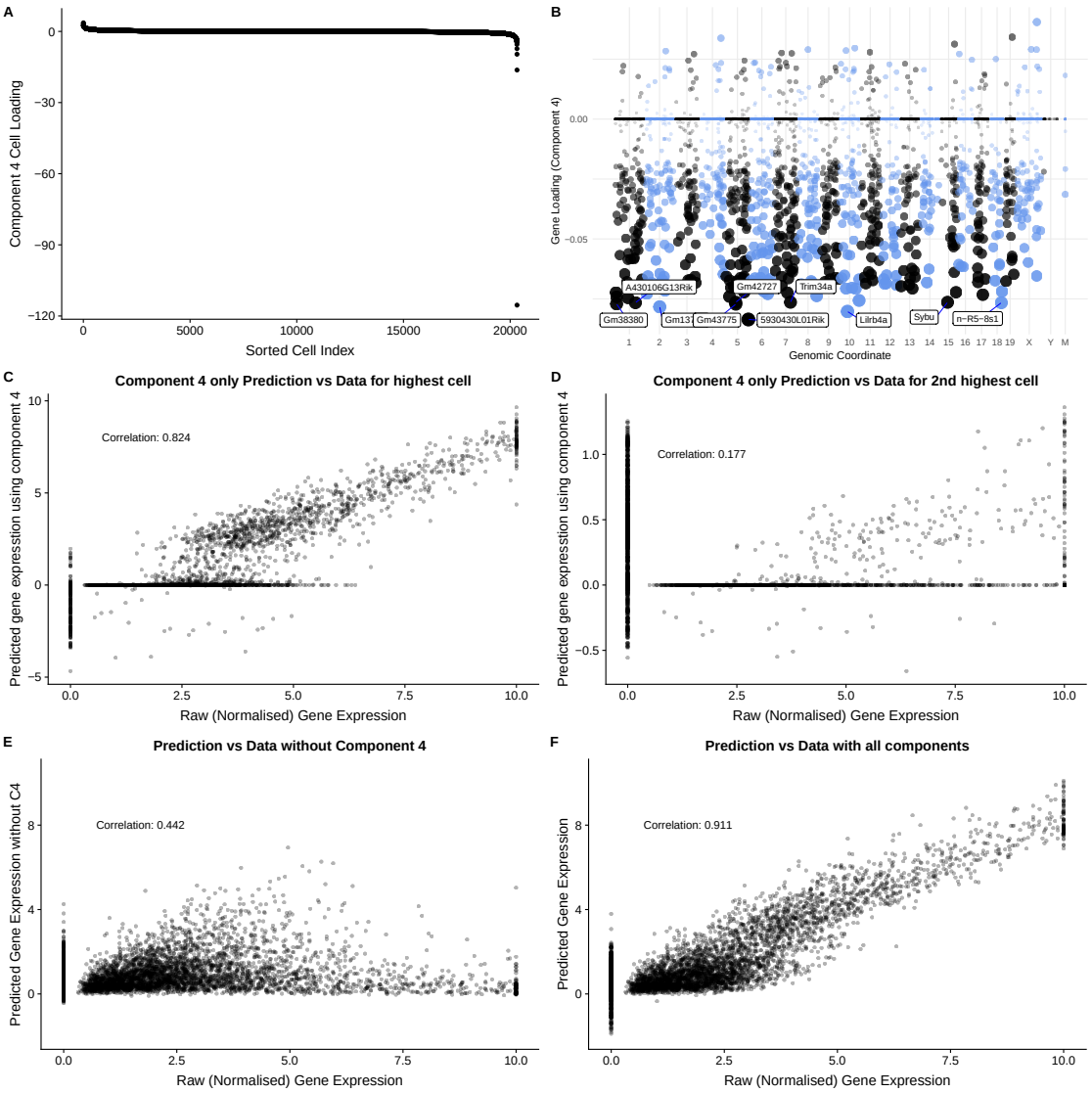


Figure 2.1: Component 4 represents a single cell. **(A)** Cell scores for component 4 ordered by value. An individual cell has outlying value. **(B)** Gene loadings for component 4. **(C)** Predicted expression using only component 4 vs Raw Normalised Gene Expression. Pearson's correlation is quoted (equal to correlation with the gene loadings). All of the genes with high expression in this cell have a high gene loading in this component. **(D)** As in C but for the cell with the second highest association. The correlation is much lower. **(E)** As in C but using all components except 4. The correlation is less than C especially for highly expressed genes. **(F)** As in C but using all components for the prediction. This is the sum of C and E

By inspecting the change in free energy as well as the change in fraction of gene

loadings with posterior inclusion probabilities (PIP) less than 0.5 we can confirm that the algorithm had converged within the 10,000 iterations for which it was run (figure 2.2). In addition the results are almost identical after 1,000 iterations (figure 2.3). Mean runtime for 10,000 iterations was 41.8 hours on a single Intel Xeon E5-2667 v4 core (SDA can be run in parallel mode). The computational complexity of SDA scales with $\mathcal{O}(NLC^2)$ where N, L and C are the number of cells, genes, and components respectively. Whilst SDA is not the fastest way to analyse single cell RNA-seq data, it may provide additional insight outweighing any gains in speed from other methods, which moreover for datasets of similar size to ours will be small relative to the time for data generation and analysis overall. As expected and in line with previous reports from Hore et al., 2016 the PIPs have a bimodal distribution (figure 2.2).

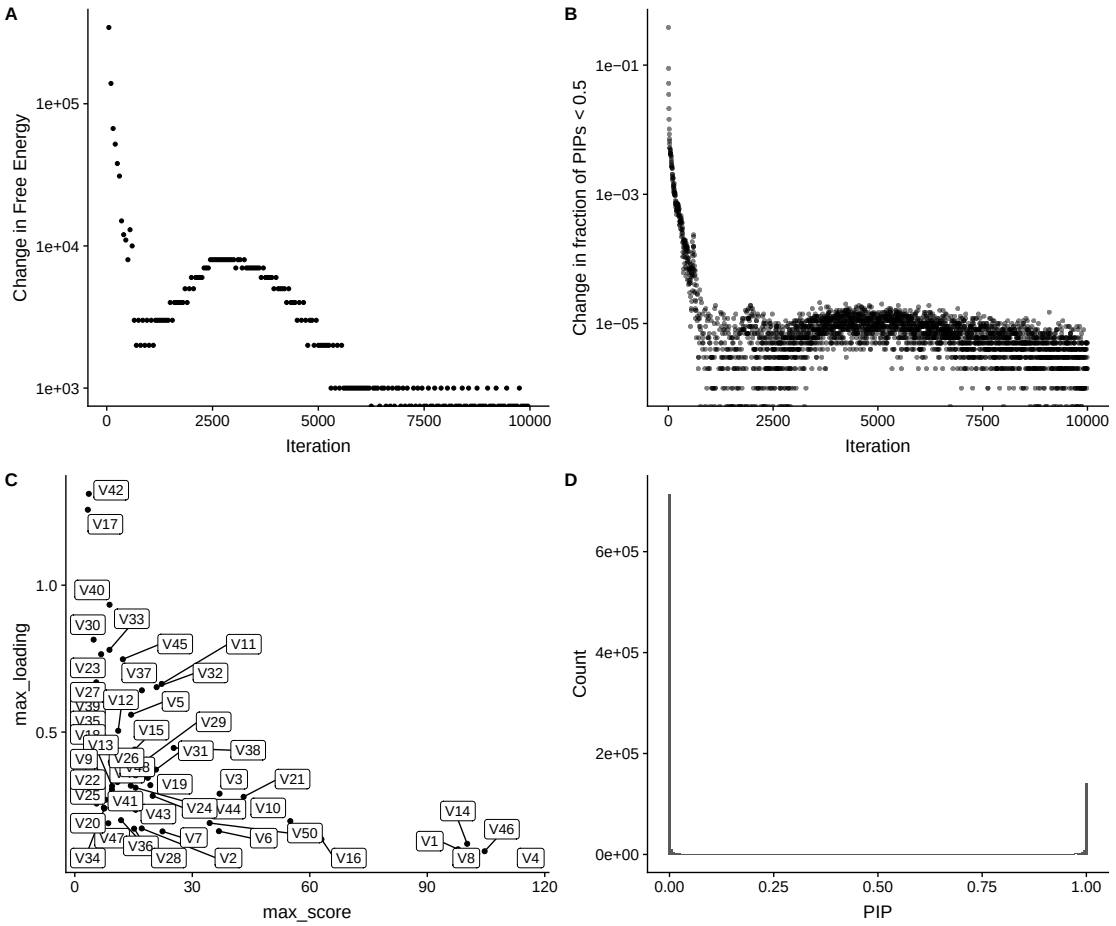


Figure 2.2: Checking convergence of SDA. **(A)** Change in free energy is often near 0 by the 10,000th iteration. **(B)** Change in fraction of PIP <0.5 is less than 0.005% **(C)** Distribution of maximum scores and loadings for each component. **(D)** Distribution of PIPs across all components, showing expected bimodal distribution.

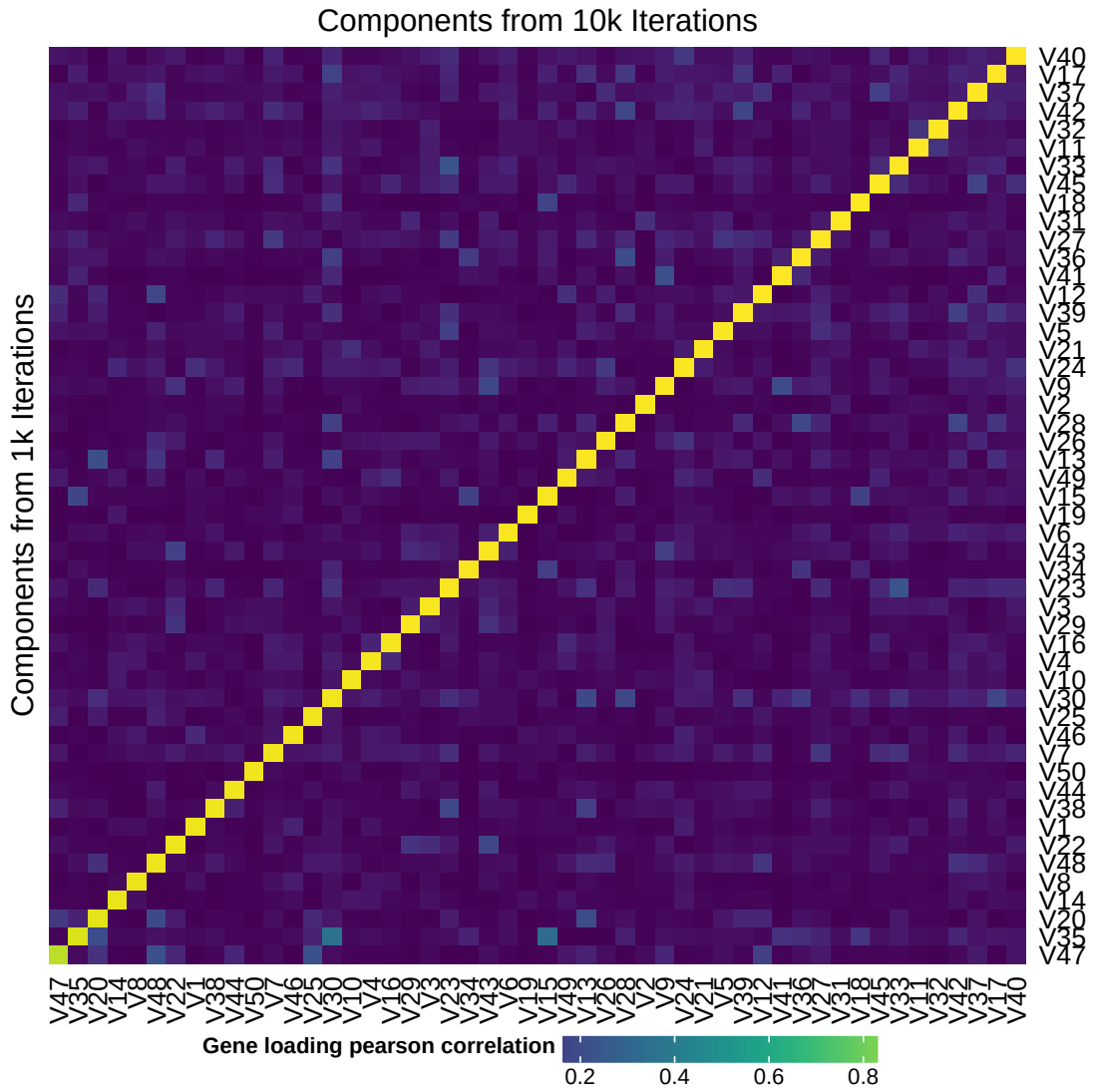


Figure 2.3: Correlation of gene loadings between components inferred using 1,000 iterations vs those using 10,000 iterations.

The gene loadings are sparse, with 79.8% of the gene loadings having a PIP of less than 0.5 (figure 2.6A,B,C). These genes have a tight distribution of loadings around 0 with a maximum absolute loading of 0.011 and 99% of the loadings lying within the range (-0.0033, 0.0036). Of the gene loadings with a PIP of greater than 0.5, the maximum absolute loading is 1.31 and the mean is 0.062.

2.5 Low Dimensional Visualisation

Despite reducing the dimensionality of the original dataset from 19,262 to 50 this is still too high to visualise the overall structure of the data. This can be achieved by performing a second non-linear reduction such as t-SNE (with perplexity parameter of 50) or UMAP (both without the usual initial PCA linear reduction step) (Maaten et al., 2008; McInnes, Healy, and Melville, 2018; Becht et al., 2018). We find a horseshoe like effect, often found when reducing datasets with an underlying 1 dimensional structure (Novembre et al., 2008; Podani et al., 2002) (figure 2.4). By looking at genes with known expression patterns from the literature we can induce that in this case the linear structure is developmental time of spermatogenesis and that the clusters in the centre contain somatic cells (figure 2.5).

To generate a pseudo-timeline (one dimensional embedding) representing this developmental time, we used a similar approach to that implemented in SCUBA (Marco et al., 2014). We iteratively fit a principal curve on the t-SNE coordinates with increasing degrees of freedom from 4 to 9 using the curve from the previous run as the initialisation (Hastie et al., 1989). Each cell was then assigned to the closest position on this curve. Somatic cells and the Hormad1 X-activated cells (component 38 score >3) were excluded during pseudotime construction, but the *Hormad1*$^{-/-}$ X-activated cells were given pseudotimes post-hoc. Somatic cells were defined by thresholding the cell scores of somatic components (if the absolute cell score of a given cell passed any of the following component thresholds: if the absolute cell score was >2 (components 26, 11, 3, 32, 45, 24); > 1.5 (component 37); > 1 (component 40); or if mt-Rnr2 expression was >3).

The temporal order of components was determined by using a weighted mean of the pseudotime values, where the weights are the cell scores of the component. In addition, only those cells with an absolute cell score of greater than two contribute to the mean.

Although the model used for inference is symmetric for positive and negative gene weightings, many identified components showed strong biases towards positive

or negative weights, consistent with expectations for identifying a group of co-activated (or co-repressed) genes enriched for specific processes (e.g. figure 2.6E&F). Likewise, the cell loadings of each component frequently highlight specific cellular subsets that localize in t-SNE space and pseudotime (figures 2.6D and 2.15)

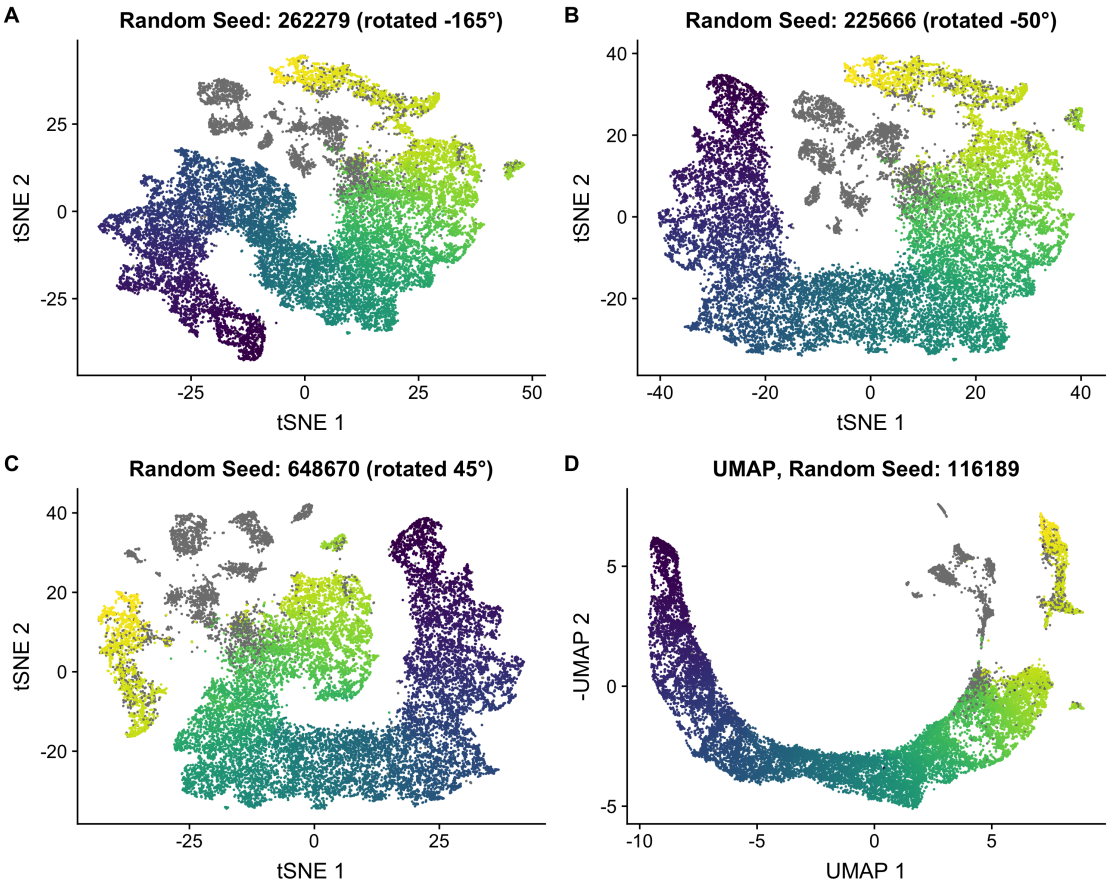


Figure 2.4: (A-C) To quantify our uncertainty in the t-SNE embedding, we performed multiple t-SNE analyses with different random seeds. The cells are coloured by the same pseudotime values used throughout. Somatic cells are coloured gray. t-SNE coordinates are rotated about the origin to aid comparison. Each was run for 1,000 iterations with a perplexity parameter of 50. **(D)** We also performed dimensionality reduction using UMAP and confirmed that it gave a pseudotime embedding consistent with t-SNE. Cells are again coloured using the same psueodtime values as A-C

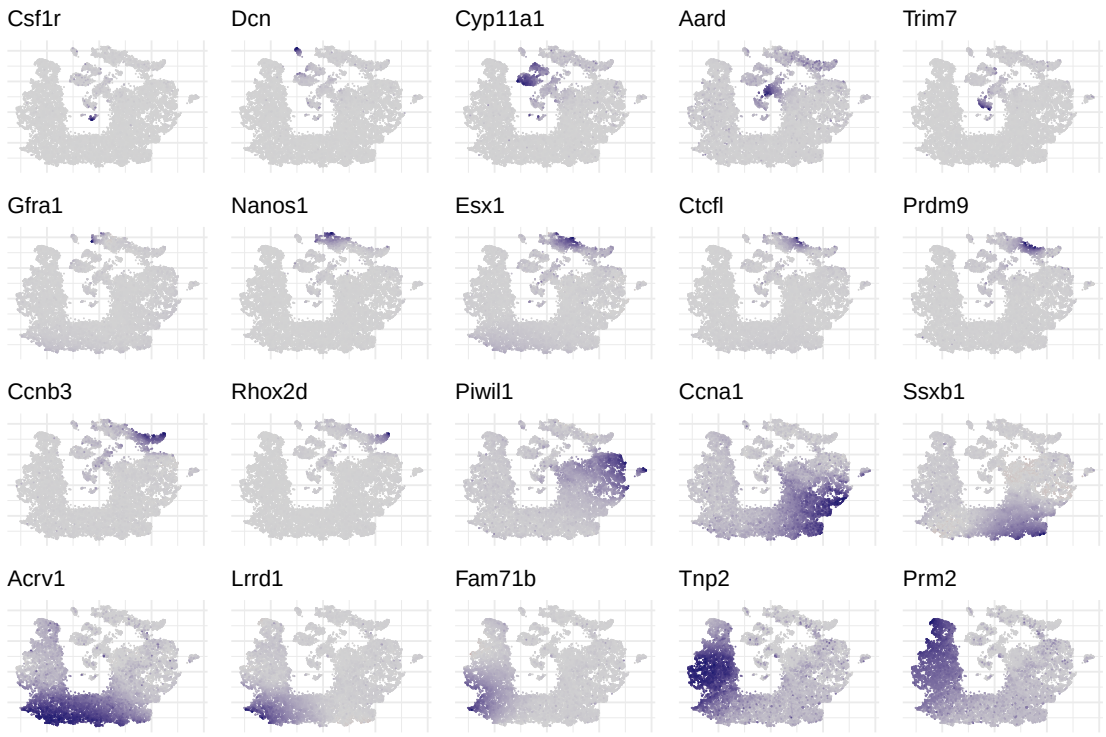


Figure 2.5: Imputed expression for genes with a range of expression patterns across cells

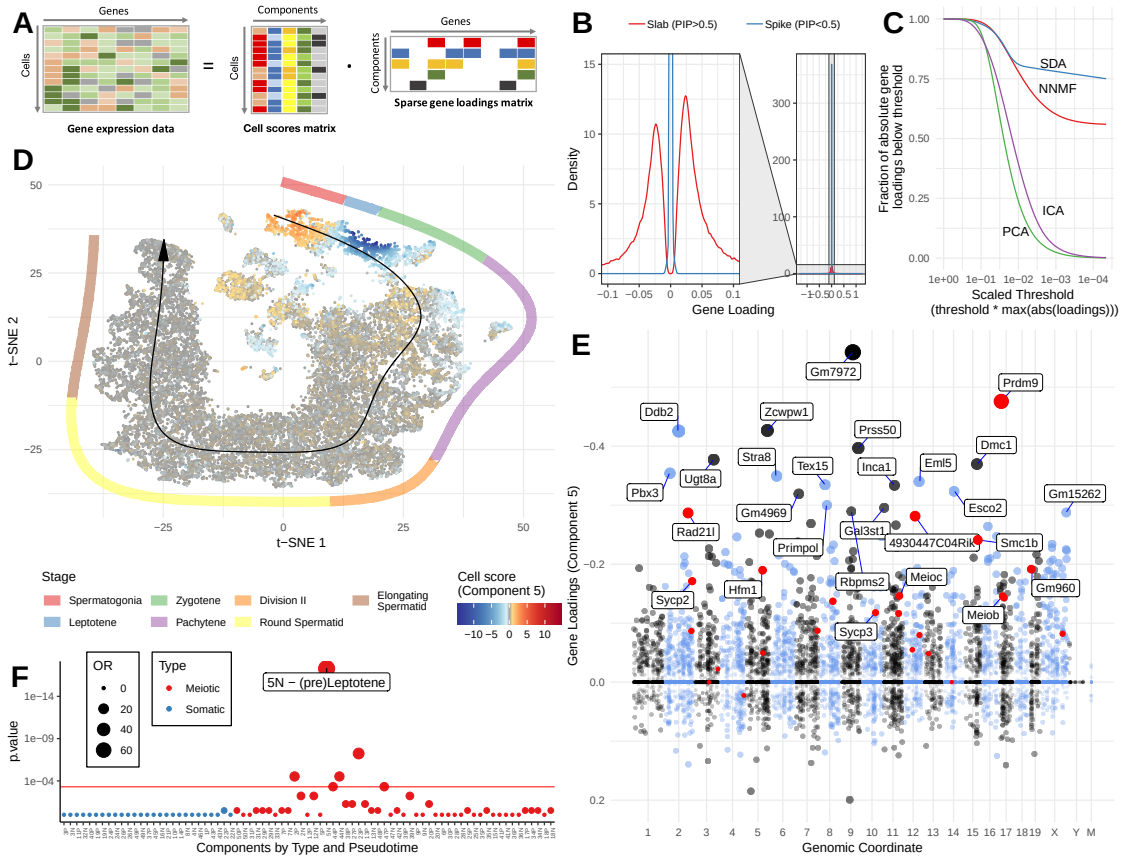


Figure 2.6: **(A)** Schematic illustrating SDA **(B)** Density of gene loadings over all components with loadings separated into genes with PIPs > 0.5 (20%) versus <0.5, indicating the sparsity of resulting gene loadings. PIP = Posterior Inclusion Probability that a gene loading is not equal to zero (i.e. not in the spike). **(C)** For each method, the fraction of all absolute gene loadings exceeding a 'no loading' sparsity threshold is shown, normalized by the maximum absolute loading across all components for that method. **(D)** t-SNE projection of the cell scores from SDA. Each point is a cell, each coloured by their loading in component 5. Black arrow: the principal curve fit to the germ cell data, corresponding to the developmental ordering of each cell progressing through spermatogenesis. The coloured segmented line shows broad staging of spermatogenesis. **(E)** Gene loadings for component 5, plotted along the genome. Red genes: GWAS hits for human recombination rate. **(F)** Enrichment for GWAS hits of human recombination rate for all components. OR: Odds Ratio. P value by FET (main text). Positive (P) and negative (N) loadings are tested separately. For one-sided components (cell score range ratio >5) the minor side is omitted. Red horizontal line: p = 0.05 after Bonferroni correction for multiple testing.

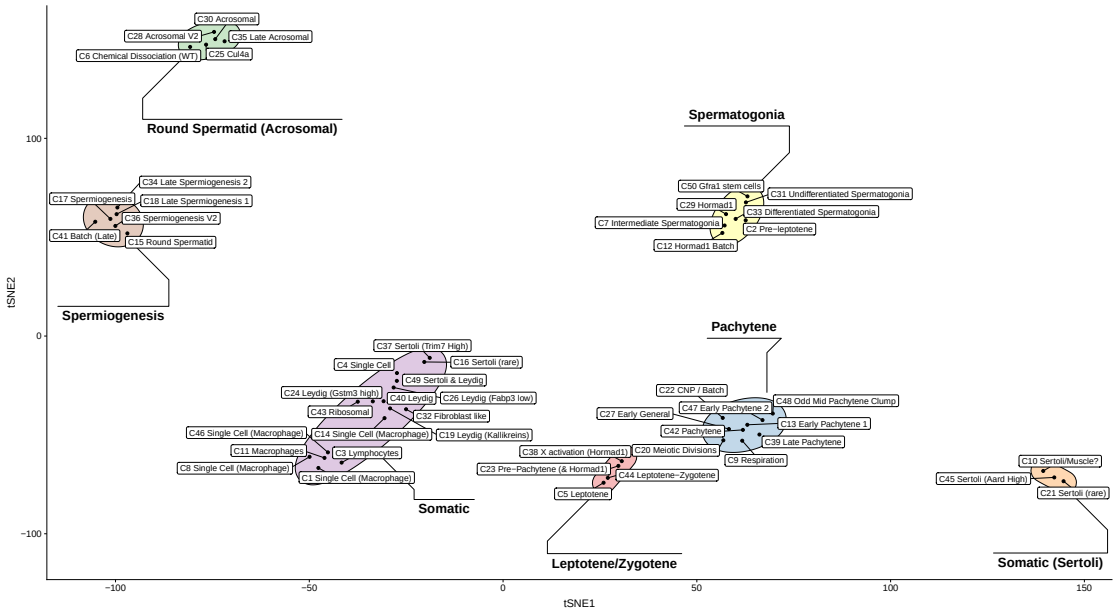


Figure 2.7: A low dimensional representation of the components shows 7 clusters.

We can perform the same dimensionality reduction on the *transposed* cell scores or gene loadings to see how the components relate to each other rather than the cells (figure 2.7). We find 7 major clusters of components and (as we later show [section 2.9]) these correspond to: 1) Spermatogonia, 2) Leptotene/Zygotene 3) Pachytene 4) Round Spermatid (Acrosomal) 5) Spermiogenesis 6) Somatic and 7) Somatic (Sertoli). We can see that our manual labelling of the components is consistent with the clusters, for example all components assigned an early label clustering together in the "Spermatogonia" cluster.

Visualisation of components through pseudotime shows that transcription during spermatogenesis can be represented as a series of overlapping components, each varying with different timescales (figure 2.8A&B). Furthermore, these components are composed of distinct gene sets enriched for distinct biological processes (figure 2.8C&D).

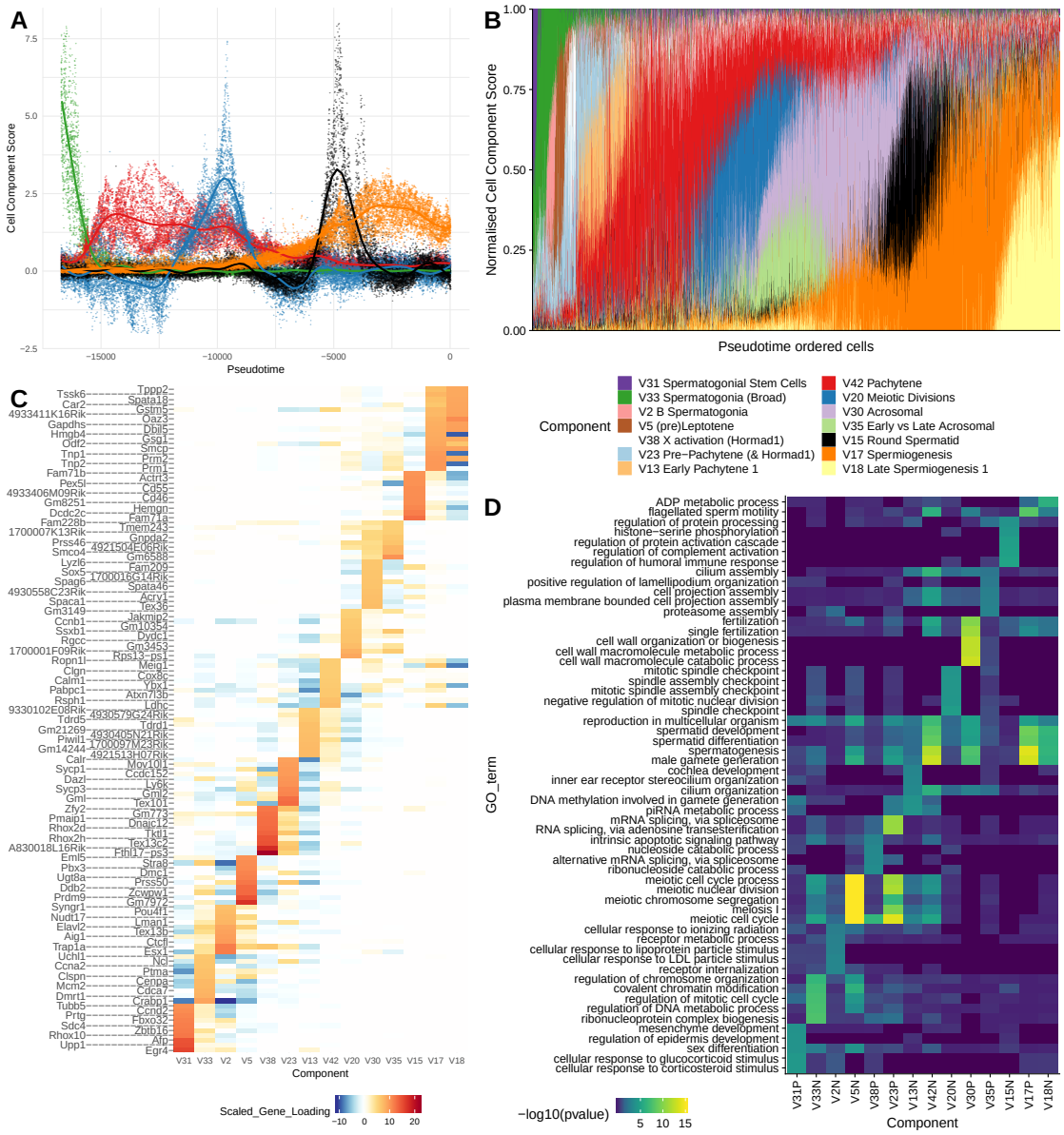


Figure 2.8: **(A)** For five example components, the cell scores for each cell are plotted through pseudotime, showing overlapping, dynamic, component activity. Component signs were defined to be mainly positive (components have arbitrary sign). Colour mappings as in panel B. **(B)** Stacked bar plot of cell component loadings for 14 germ components sorted by cell pseudotime. Each column corresponds to an individual cell and the total positive component loadings for each are normalized to one after flipping components to be mainly positive. **(C)** Shown are the top 10 gene loadings for each of the components in (B) represented as a heatmap. Most genes have strong loading on only one component. **(D)** Gene ontology enrichment analysis for biological processes in the top 250 genes for each component.

2.6 Imputation

In addition to identifying soft clusters and their marker genes, In the process of fitting the SDA model we have also effectively imputed the very sparse and noisy data. Multiplying out the cell scores and gene loadings matrix generates a matrix of the same dimensions as the original data but populated with the models predicted underlying values. Encouragingly, SDA imputation is able to estimate expression of individual genes even when in many cells zero reads are observed (figure 2.9A). In addition SDA expression estimates have fewer outlying high values outside the genes main expression window. The destructive nature of the single cell RNAseq protocol means that it is not possible to determine the true expression vector for an individual cell. In order to determine if imputation is actually improving our estimates we use cross-validation. Specifically, we randomly assign each read to either a training or test set, predict gene expression based on the training set (using SDA, or another method for example the dedicated imputation method MAGIC from van Dijk et al., 2018) and then evaluate our ability to rank gene expression using the test set (methods section 2.12.4). SDA imputation outperforms approaches using the raw data for essentially all cells in the test data (figure 2.9B&C).

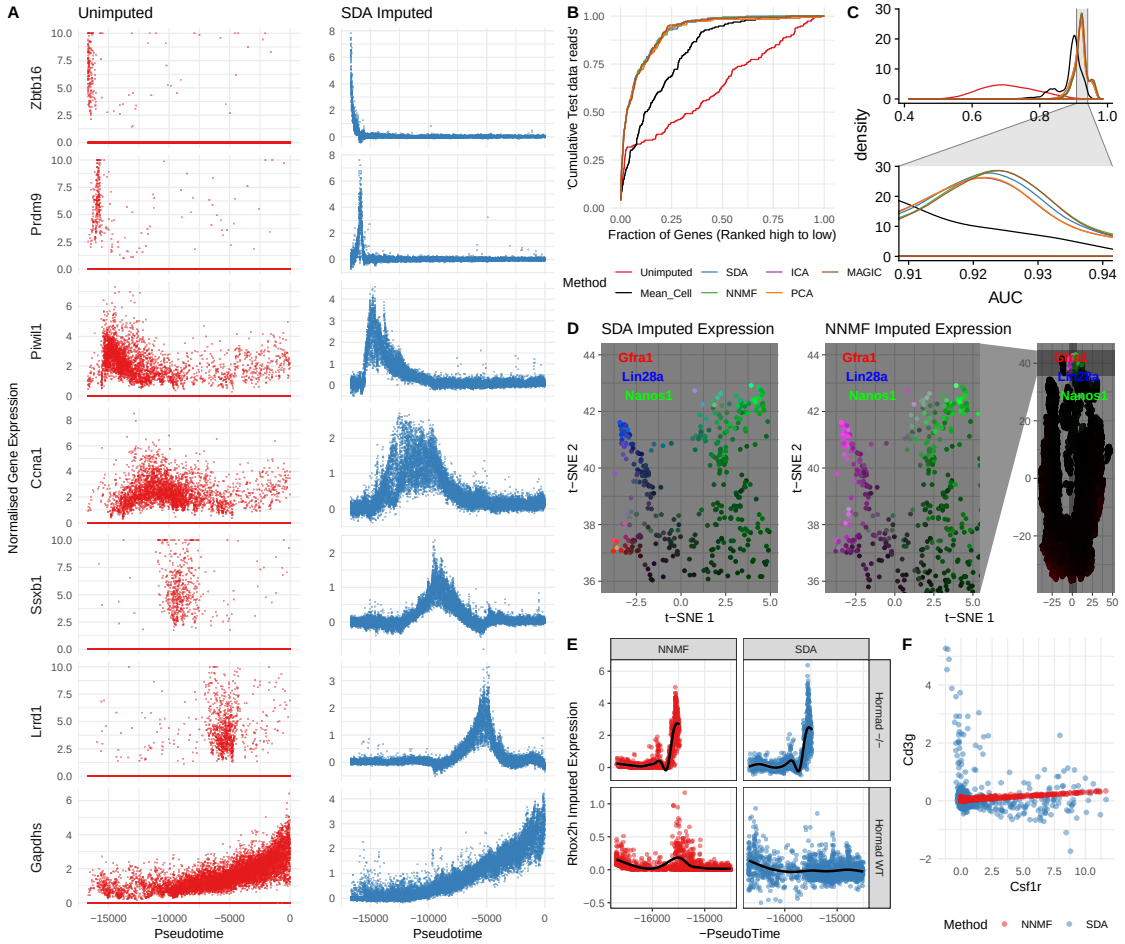


Figure 2.9: **(A)** For seven genes expressed at various stages of spermatogenesis unimputed normalised gene expression is shown alongside SDA imputed expression. **(B)** Prediction accuracy of seven predictors of gene expression trained on test data for an example individual cell. 'Unimputed' uses the training data directly, 'Mean Cell' uses the mean across all cells, matrix factorisation approaches SDA, PCA, ICA, NNMF, and a dedicated imputation approach, MAGIC. **(C)** Comparison of AUCs (Area under the curve) for all cells using various methods (same colour scheme as part B). **(D)** Zoomed versions of the t-SNE projection (with full t-SNE for context): cells are coloured by expression using a three channel ternary colour scheme with the amount of blue, green, red representing the respective expression levels of *Lin28a*, *Nanos1*, and *Gfra1*. **(E)** Imputed expression of X chromosomal gene *Rhox2h* from either the SDA or NNMF decomposition, split into cells we know to be either WT or *Hormad1*$^{-/-}$ genotype. **(F)** Imputed expression of *Cd3g* and *Csf1r* for both NNMF and SDA.

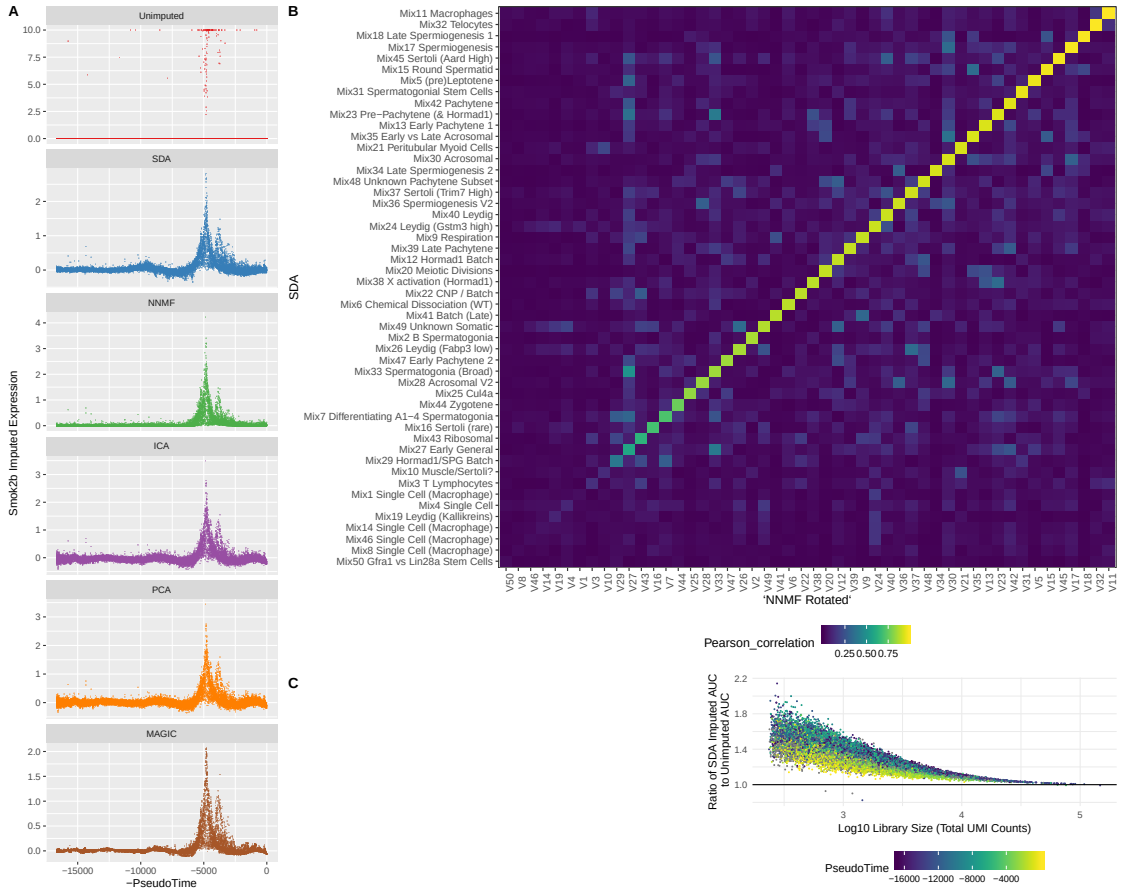


Figure 2.10: (A) Imputed expression of an example gene (*Smok2b*) for different methods, to illustrate the similar predictions as shown in figure 2.9B and 2.9C. **(B)** Overall, NNMF infers similar components to SDA. The heatmap shows Pearson correlations between different pairs of gene loading vectors from SDA and NNMF (with procrustes rotation applied, methods section 2.12.4). **(C)** The fold improvement in AUC when comparing SDA imputation to the unimputed data, plotted as a function of cell library size.

While providing the most sparse representation (figure 2.6C), SDA still imputes equally well, compared to NNMF, other matrix factorisations, and MAGIC (van Dijk et al., 2018) (figure 2.9C & 2.10A). Some cells gain more from imputation than others, the cells that gain the most are those with low library size (figure 2.10C).

Furthermore, by comparing the genes with the lowest correlation between NNMF and SDA imputed expression, we found that SDA provides additional biological insights for the same number of components (figure 2.9D,E,F & 2.10B). SDA infers multiple components for undifferentiated spermatogonia whereas NNMF only infers a single component resulting in imputed expression of *Gfra1* and *Lin28a* in the same cells (figure 2.9D and figure 2.10B - no correlation for SDA component 50 *Gfra1*

Stem Cells). NNMF predicts a peak in the expression of X-linked gene *Rhox2h* even in WT cells, in which X chromosome activation due to Hormad1 KO does not occur, suggesting underfitting of the data with 50 components (figure 2.9E). NNMF does not predict high expression of the adaptive immune cell marker *Cd3g* (T-cell surface glycoprotein CD3 gamma chain), and when it predicts any expression it increases linearly with the innate immune cell marker *Csf1r* (Macrophage Colony-Stimulating Factor 1 Receptor, or *Cd115*) (figure 2.9F). In contrast, SDA correctly predicts that *Cd3g* and *Csf1r* are not coexpressed in the same cells (see also figure 2.10B, no correlation for the SDA component 3 Lymphocytes).

In addition to obviating the need for further clustering and differential expression analyses, an advantage of using matrix factorisation for imputation is the much smaller memory footprint required to store the results: on our dataset MAGIC produces 2.9 Gb of imputed data, whereas the SDA matrices are just 18 Mb (12.6 Mb when loadings with PIP <0.5 are set to 0).

2.7 Mutant Genotypes

Given we combined both wild-type and mutant cells in our analysis it is important to check that the components representing wild type processes are not unduly affected by this mixing. In order to quantify the robustness of our conclusions to this decision to combine mutant and wild-type strains, we performed a separate SDA analysis using just wild-type cells (the 'WT' analysis) and compared the results.

Factor analyses naturally have a degree of unidentifiability, whereas in SDA the sparsity prior helps to make the model identifiable across different seeds. For different data however, the gene loadings matrix may be a rotated version in which components are linearly split or combined. A procrustean rotation can align two matrices (here gene loadings) onto an equivalent set of axes. Therefore, for fair comparison across SDA runs with different datasets, we compared the correlation of the gene loadings after procrustean rotation which tells us if the components inferred jointly capture the same information.

We found that most components generated from an SDA analysis of only wild-type data were also observed in the joint analysis of wild-type and mutant data. The most important WT components (those with high total cell score) have high correlations with components in the mixed SDA run (figure 2.11). Some components such as "Mix38 X activation" are not represented in the WT decomposition because they represent mutant-specific processes. Other components such as "Mix44 Leptotene-Zygotene" do not appear as these cells are numerous only in mutant samples which lack the more abundant later stages of cells.

By visualising the rotation matrix from the procrustean analysis we can see which components have been rotated. For example "Mix17 spermiogenesis" is a linear combination of WT component 46 and 7 (figure 2.12). We can also visualise the correlations directly as in figure 2.13.

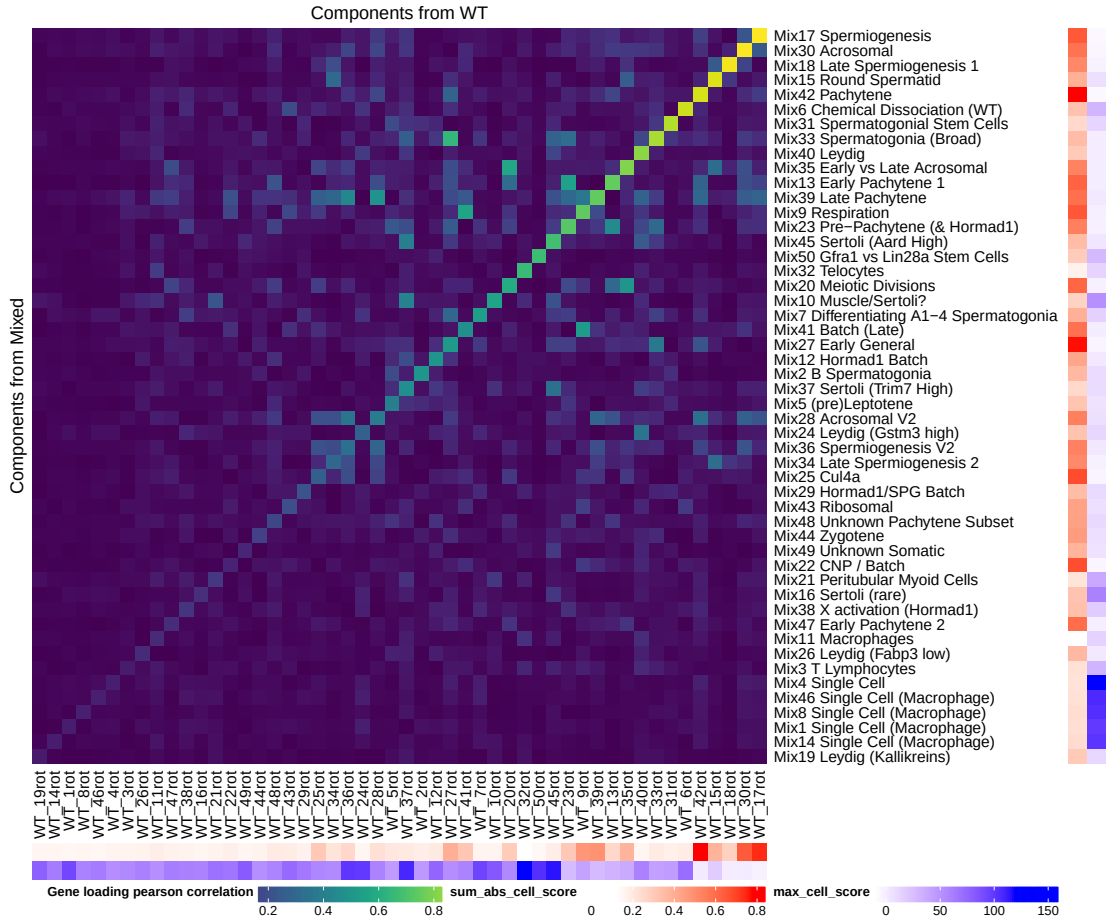


Figure 2.11: Here we show as a heatmap the pearson correlation of component gene loadings between a procrustean rotation of the WT gene loadings and the Mixed SDA gene loadings. The 'sum abs cell score' annotation shows the sum of the absolute cell scores for that component (larger number indicates a more important component). The 'max cell score' annotation indicates the maximum cell score for each component (a larger maximum indicates potential over-fitting to a single/small number of cells).

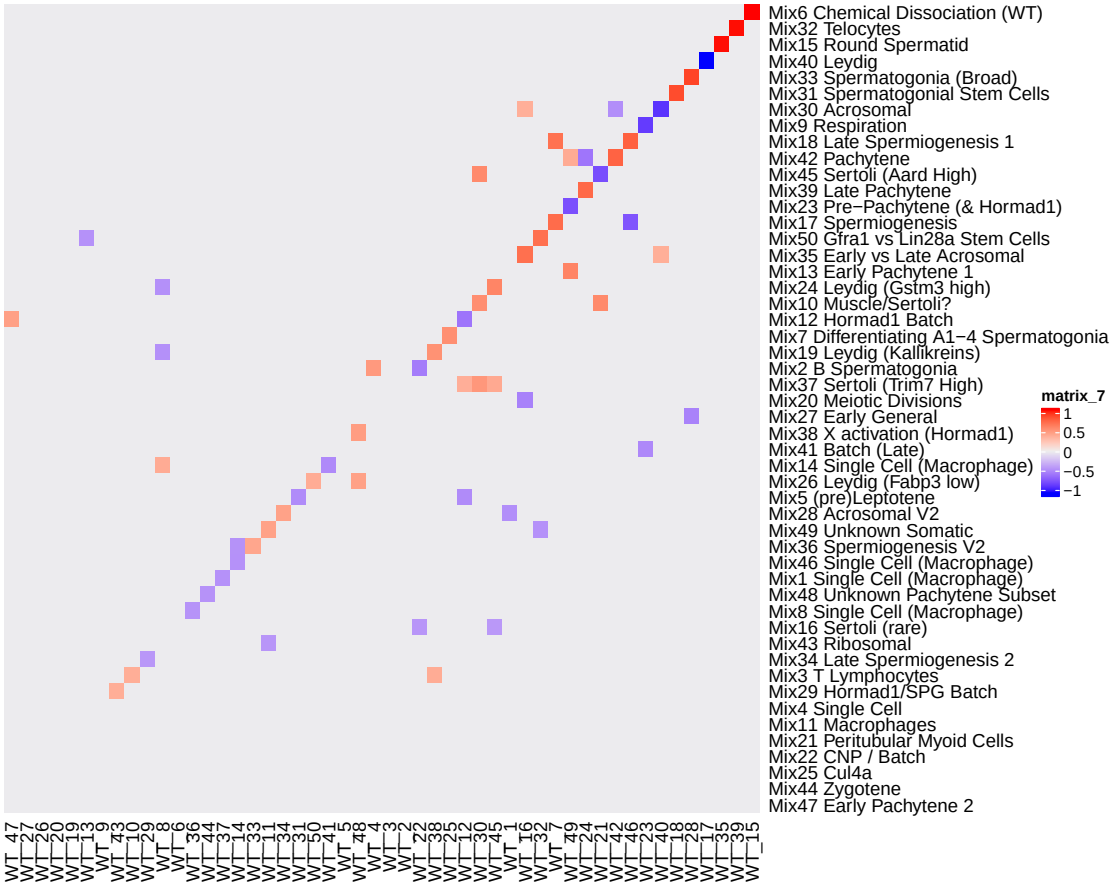


Figure 2.12: This heatmap shows the procrustean rotation matrix (with absolute values < 0.35 rounded down to 0).

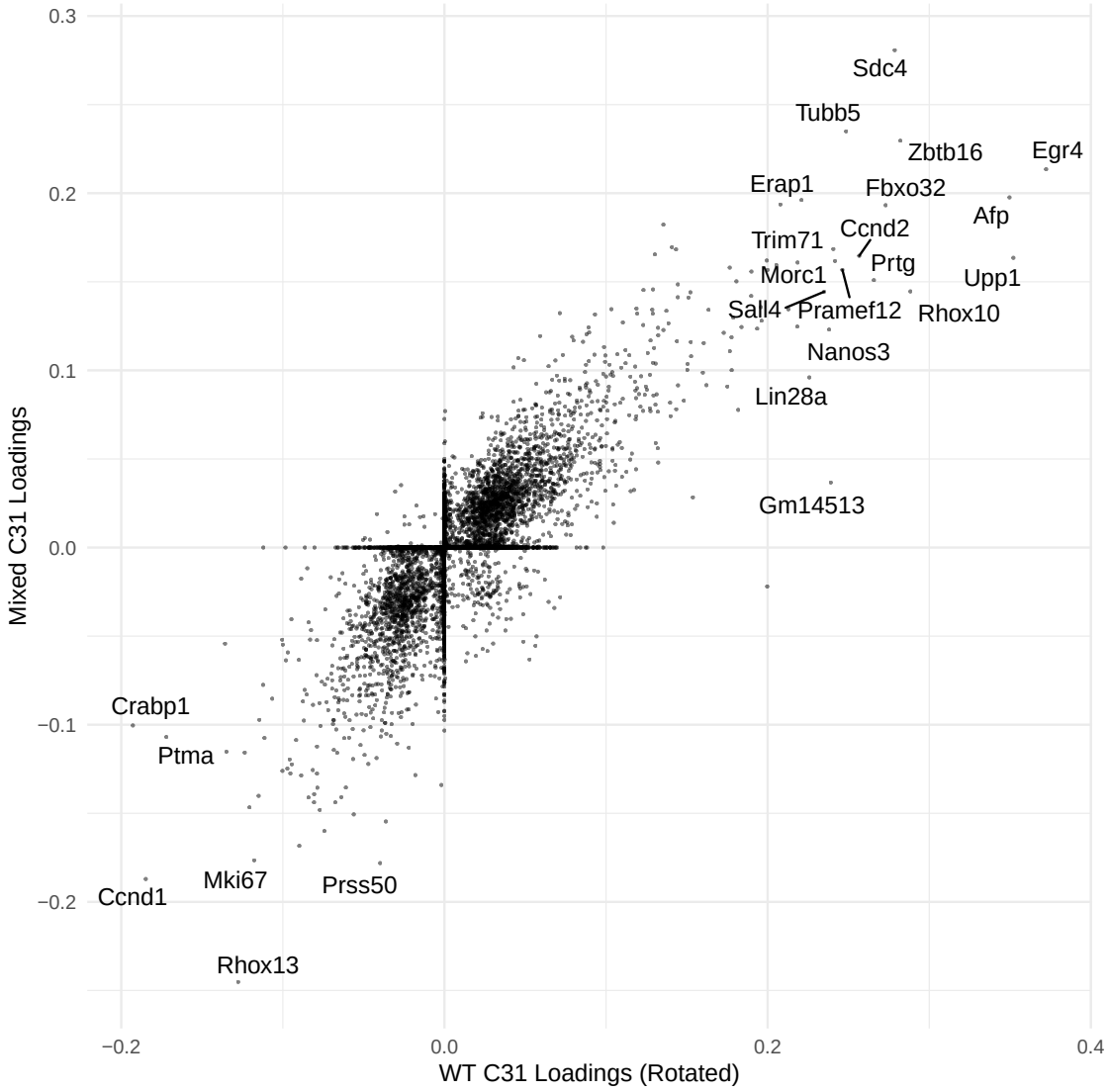


Figure 2.13: An example scatterplot comparing the gene loadings for one cognate SDA component (C31) between WT and Mixed SDA runs. The correlation is high.

2.8 Potential Artefactual Components

Given we combined multiple experimental batches it is important to correct for batch effects and SDA provides a simple approach as some inferred components will capture these effects and can be removed.

Component 22 gene loadings are highly enriched in ribosomal proteins (50 of the top 100 loadings are non pseudogene ribosomal genes - of which there are 99, figure 2.14A). However unlike ribosomal component 43, the cell scores for component 22 are associated with whether the sequencing was done on MiSeq or HiSeq Illumina machines and so this component is likely a batch effect (figure 2.14B).

Component 12 has high cell loadings for *Hormad1*$^{-/-}$ cells, but these loadings are largely positive for the first two libraries, and negative for the last library (figure 2.14C,D).

Component 9 contains 41 out of 57 of the respiratory complex I, III, and IV genes (complex II is an alternative pathway) (figure 2.14E,F, $p = 7.4 \times 10^{-53}$, OR = 104 [95% CI: 62.1,Inf] FET). It's not clear whether this components' activity reflects real meiotic biology or potentially reflects cell stress induced by experimental processing (still a real biological process but not one that exists in vivo).

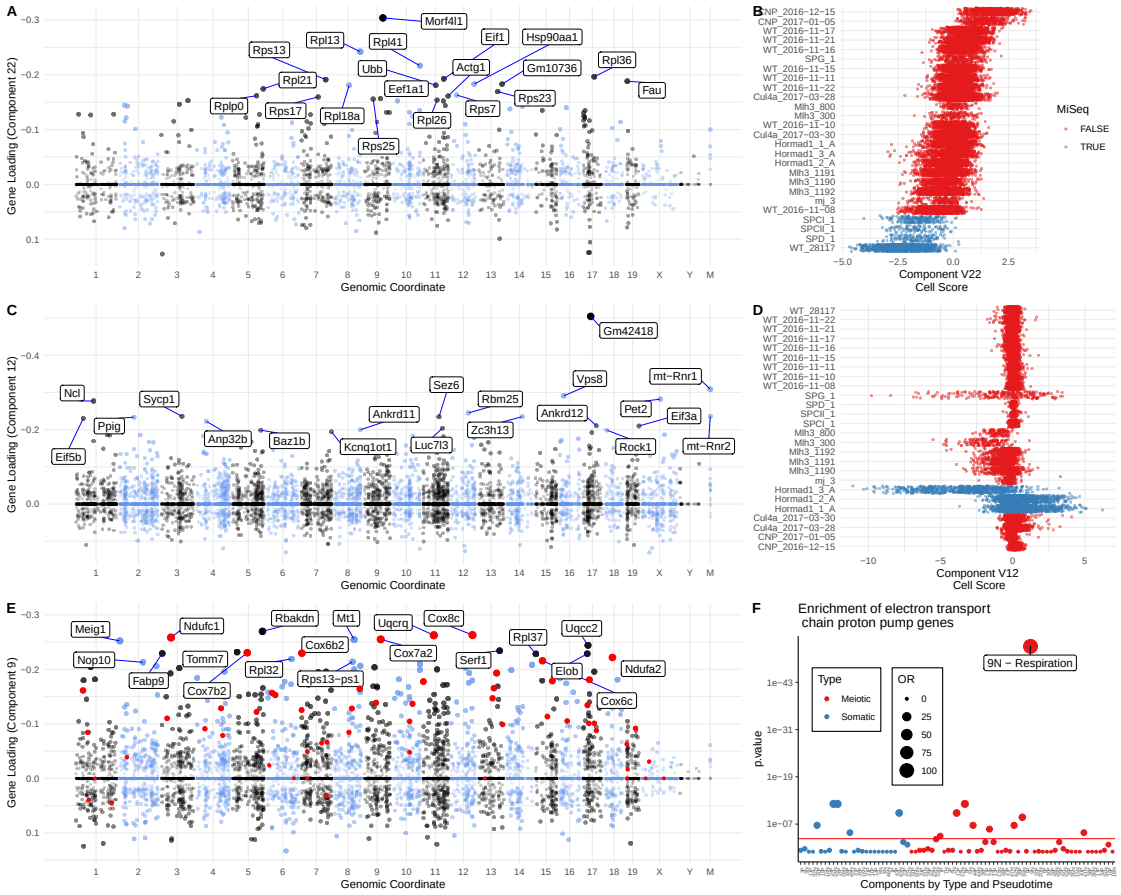


Figure 2.14: (A & B) Gene loadings and cell scores for component 22. **(C and D)** Gene loadings and cell scores for component 12. **(E)** Gene loadings for component 9. Genes that are (non-assembly) components of the electron transport chain proton pump genes are highlighted in red. This gene set is defined by the genes that match the regex 'Uqc|Cox|Ndu'. Pseudogenes, and genes with 'assembly|like' in their name were excluded. **(F)** Enrichment of genes that are (non-assembly) components of the electron transport chain proton pump genes in the top 500 genes from each component.

2.9 Components

Assigning a cytologically defined stage to each component requires considerable manual work due to lack of any previous high resolution annotations. There are a number of issues that complicate assignment. Firstly as we have seen components activities are gradual and so there may be no well defined set of cells or start and end to a component. Components can be mixed sign (positive/negative), and one sign may be active in a different stage to the other. Stages themselves are not well defined in the literature with multiple definitions being used by different groups.

Due to translational regulation the transcription of genes may not correspond to the detection of the protein, hence using commonly performed antibody based immunofluorescence can be misleading and so, if available, mRNA detection methods should be used for reference. Many studies only check protein expression, and often only by western blot from whole tissues to check testis specificity rather than spermatogenic stage. Even when mRNA is quantified developmentally the time points are often so widely spaced as to only permit vague conclusions about when expression starts. In addition, any literature references from human, or even rat, can have different expression patterns than mouse and so this evidence must be used with caution. Nonetheless, using a combination of evidence types from multiple genes for each component increases our confidence and precision enough such that we assigned each component to a stage(s) or cell type(s) most consistent with the available data.

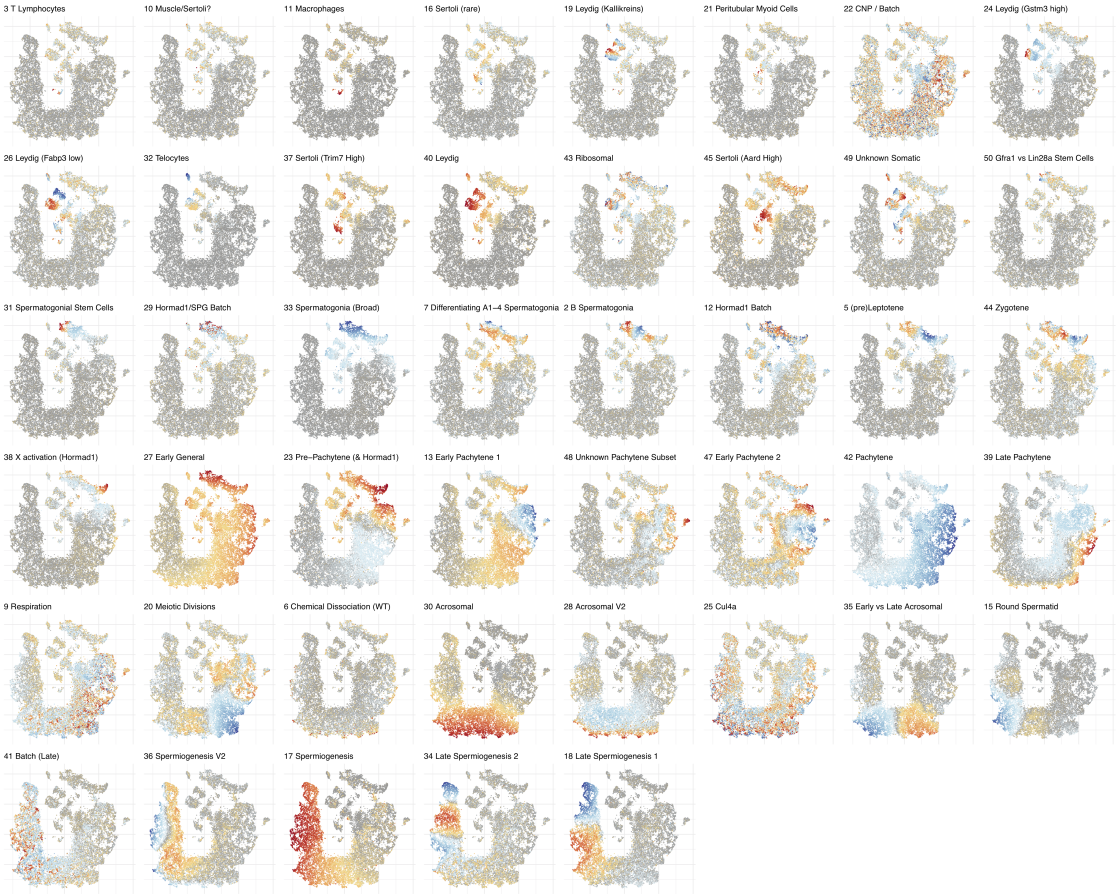


Figure 2.15: Cell scores as coloured t-SNE plots, excluding single cell components

2.9.1 Telocytes

Recently telocytes were found to be present in the mammalian (human) testis (Marini et al., 2018; Kuroda et al., 2004). We find one component (#32) matching the known combination of markers for telocytes (*Cd34* and *Pdgfra* positive, *Kit*, *Pecam* and *Acta2*/α-SMA negative). We also find many other genes are specific to this cluster (*Dcn*, *Col1a2*, *Col1a2*, *Col3a1*, *Col6a1*, *Col4a4*, *Col4a1*, *Col1a1*, *Lamb1*, *Lama2*, *Lamb2*, *Gsn*, *Adamts5*, *Mgp*), representing potential novel markers for this cell type in the testis. This component is highly enriched for the GO term "extracellular matrix organization" ($q = 4.3 \times 10^{-16}$, OR = 10.3). These cell's expression pattern matches what was described as an unknown mesenchymal cell population by Green et al., 2018 with expression of *Tcf21*, *Arx*, *Vim*, and *Col1A1*, but not *Acta2* (α-SMA), or *Cyp17a1*.

2.9.2 Leydig Cells

Component 40 broadly marks all Leydig cells and has high gene loadings for all the major genes required for testosterone production including *Star*, *Cyp11a1*, *Hsd3b*, *Cyp17a1*, and *Hsd17b* (figure 2.19F, Stojkov et al., 2013) in addition to other markers of Leydig cells including *Insl3* and *Ptgds* (Balvers et al., 1998; P. J. Baker et al., 2001). This component is most highly enriched for the GO term "steroid metabolic process" ($q = 5.3 \times 10^{-23}$, OR = 11.3)

Component 26 splits Leydig cells into two main groups, distinguished by the expression of *Fabp3*, *Gstm1*, *Gsta3*, and *Ass1*.

Component 24 is a subtype of the *Fabp3* expressing C26 positive cells, which has high expression of *Gstm3*, *Mt3*, *Thrsp*, and *Jak3*, but also a number of pseudogene versions of genes that are normally expressed leydig genes: *Gstm2-ps1*, *Gm6665* (another *Gstm2* pseudogene), *Fabp3-ps1*, *Gm8834* (*Gstm3* pseudogene), *Gm5096* (*Bhmt* pseudogene), *Gm6977* (*Fth1* pseudogene), *Gm7049* (*Me1* pseudogene).

Component 19 marks cells in both major groups and is distinguished by high expression of a group of endopeptidases encoded in a 300kb locus on chromosome 7 including *Klk1*, *Klk1b1*, *Klk1b21*, *Klk1b22*, *Klk1b24* and *Klk1b27*. Some of these

genes are known to be specifically expressed in Leydig cells and may be involved in remodelling of the extracellular matrix (Sanz et al., 2013; Matsui et al., 2000; Matsui et al., 2001; Matsui et al., 2005).

2.9.3 Sertoli Cells

Component 37 is active in cells expressing *Wipf3* (aka CR16) and *Ncoa2* (aka TIF2) both of which are required for fertility and have Sertoli restricted expression (Suetsugu et al., 2007; Gehin et al., 2002)). In addition *Nxf3* is specifically expressed in Sertoli cells but not required for spermatogenesis (J. Zhou et al., 2011).

Component 45 is also likely to be active in Sertoli cells, but merges with the pachytene cells, very speculatively representing phagocytosis of apoptosing cells.

2.9.4 Macrophages

We find one component (#11) representing macrophage cells. This component has high loadings for *Csf1r* (macrophage colony-stimulating factor receptor, aka CD115), *Cd163* (macrophage scavenger receptor), *Cd68*, *Adgre1* (aka F4/80), *Itgam* (aka CD11b), *Mrc1*, *Cx3cr1*, *Fcgr3*, and the complement genes *C1qa*, *C1qb*, and *C1qc* (Mossadegh-Keller et al., 2017; Fabriek et al., 2005; Sasmono et al., 2012). Macrophages are professional antigen-presenting cells and the highest enriched GO term for this component is "antigen processing and presentation of peptide antigen" ($q = 1.4 \times 10^{-19}$, OR = 9.5) driven by genes including the MHC I and II genes *H2-D1*, *H2-Aa*, *H2-Ab1*, *H2-Eb1*, *H2-K1*, *H2-DMa*, *H2-T23)*, *B2m*, and the other immune related genes such as *Cd74*, *Fcgr3*, *Fcgr2b*, *Fcer1g*.

Immunohistochemical staining of ADGRE1 in *Hormad*$^{-/-}$ and *Mlh3*$^{-/-}$ but not wild type mice found that macrophages were present within the tubule, which is typically regarded as immune privileged partly due to the blood-testis barrier formed by Sertoli cell tight junctions (Fijak et al., 2006). Macrophages have previously been reported within the adluminal compartment, although always in the context of testicular defects (Frungieri et al., 2002; Goluža et al., 2014). Intriguingly component 11 is also specifically and strongly enriched for GWAS hits

for Alzheimer's ($p = 1.3 \times 10^{-11}$, OR = 65.9 by FET, figure 2.19E) and these cells are also marked by the Alzheimer's gene *Apoe* in *Mlh3*$^{-/-}$ (25/707 tubules, 3.5%, $p < 2.2 \times 10^{-16}$) and *Hormad*$^{-/-}$ (98/1756 tubules, 5.6%, $p < 2.2 \times 10^{-16}$) but not wild type tubules (0/2496 tubules). Co-staining of ADGRE1 with an antibody for activated CASPASE-3, a marker of apoptosis, failed to identify any double positive cells, excluding the possibility that intratubular ADGRE1 protein expression was somehow an artifact of an apoptotic cell population. The mechanisms by which macrophages transit the blood-testis barrier, and the corresponding cues for migration, await further investigation.

2.9.5 Lymphocytes

Component 3 is highly enriched for genes involved in T cell activation ($q = 1.4 \times 10^{-19}$, OR = 7.6) and has high loadings for (T-cell) lymphocyte genes including *Ptprc* (aka CD45, leukocyte common antigen), *Cd2* (aka T-cell surface antigen) (Murray et al., 2011; Murphy et al., 2012), *Ms4a4b* (high protein expression in T and NK but not B cells - H. Xu et al., 2010), and the CD3 T-cell receptor complex genes: *Cd3g*, *Cd3d*, *Cd3e*, *Trbc2*, *Trac* (Call et al., 2002).

2.9.6 Peritubular Myoid

Component 10 is active in a small number of cells, likely to be peritubular myoid cells due to co-expression of *Cnn3*, *Edn1* (Endothelin), *Myh11* (Smooth muscle myosin heavy chain), and *Acta2* (Smooth muscle actin) (Mayerhofer, 2013). The most significant GO term is "muscle tissue development" ($q = 8.8 \times 10^{-8}$, OR = 4.97).

2.9.7 Spermatogonia

Five components correspond to processes in spermatogonia. Component 31 represents undifferentiated spermatogonia expressing *Zbtb16* (aka *Plzf*) (Buaas et al., 2004) and *Foxo1* (Goertz et al., 2011), while component 50 splits these cells into two subpopulations, the first expressing *Glis3* (H. S. Kang et al., 2016) and *Gfra1* - shown to be expressed in a subset of A_s and required for A_s self renewal by acting

as a co-receptor for GNDF (Meng et al., 2000; Z. He et al., 2007). The second subpopulation expresses *Nanos3* (Suzuki et al., 2009), *Lin28a* (K. Zheng et al., 2009) and *Foxf1*. Component 7 likely represents A_{1-4} spermatogonia as they express c-Kit which is expressed in differentiated (not undifferentiated) spermatogonia (Manova et al., 1990; Schrans-Stassen et al., 1999) and is required for the A1 to AIn divisions (Yoshinaga et al., 1991). These cells also express *Stra8* which is expressed from type A spermatogonia through to preleptotene cells in response to retinoic acid (Oulad-Abdelghani et al., 1996; Q. Zhou et al., 2008; Endo et al., 2015) and is required for premeiotic DNA replication (Baltus et al., 2006). We find these cells are also well marked by *Plppr3*, *Sertad4*, *Met*, *Jade2*, *Nanos1*, *Glis1* and *Glis2*. Component 2 includes *Ctcfl*, *Pou4f1*, and *Esx1* - likely representing intermediate and type B spermatogonia (Sleutels et al., 2012; Budhram-Mahadeo et al., 2001; Maezawa et al., 2018; Y. Li et al., 1997; Branford et al., 1997). Component 33 is a broad component covering all the stages of spermatogonia.

2.9.8 (pre)Leptotene & Zygotene

Component 5 includes many genes required for the creation and repair of meiotic double strand breaks. This includes *Prdm9* itself; components of the meiotic cohesin complex *Rad21l*, *Smc1b*, *Smc3*, *Stag3* and *Esco2* (Rankin, 2015); components of the telomere tethering complex *Terb1*, *Terb2*, *Spdya*, and *Sun1* (Ding et al., 2007; Tu et al., 2017; Y. Wang et al., 2019); genes involved in creating DSBs *Mei1*, *Ccdc36* (*Iho1*), *Spo11* partner *Top6bl* (*Gm960*), and regulator *Atm* (Lukaszewicz et al., 2018; Reinholdt et al., 2005; Robert et al., 2016; Stanzione et al., 2016; Vrielynck et al., 2016); proteins required for the creation and processing of the ssDNA intermediates and their regulators: *Mcm8*, *Dmc1*, *Rad51*, *Rad51ap2*, *Atr*, *Brca2*, *Tex15*, *Meilb2* (*Hsf2bp*), *Meiob*, and *Spata22* (Brown, Grubb, et al., 2015; Brown et al., 2014; Dai et al., 2017; Kovalenko et al., 2006; K. Y. Lee et al., 2015; Martinez et al., 2016; Pacheco et al., 2018; Ribeiro et al., 2018; Widger et al., 2018; Y. Xu et al., 2017; F. Yang et al., 2008; J. Zhang et al., 2019); class I crossover (ZMM group) proteins *Shoc1* (*Zip2* orthologue), *Tex11* (*Zip4* orthologue), *Msh5*, *Hfm1* (*Mer3*

orthologue) and regulator *Brip1* (*FancJ*) (Adelman et al., 2008; Guiraldelli et al., 2018; Guiraldelli et al., 2013; Rakshambikai et al., 2013; X. Sun et al., 2016); as well as components of the synaptonemal complex *Sycp1*, *Sycp2*, *Sycp3*, *Syce2*, *Syce3*, *Tex12*, and *Six6os1* (*4930447C04Rik*) (Gómez-H et al., 2016; Syrjänen et al., 2014).

This component is also highly enriched for GWAS hits of recombination rate in Icelandic humans (Halldorsson et al., 2019). Of the 24 significant GWAS loci identified with confidently associated causal genes, more than half (13) rank within the top 300 genes of this component, and almost all (20) rank within the top 1300 genes ($p = 5.2 \times 10^{-18}$, OR = 77.8 [95% CI:36.4,Inf] and $p = 2.4 \times 10^{-20}$, OR = 70.1 [95% CI:26.9,Inf] respectively by fisher's exact test [FET]). One of the hits, *Msh4*, is not ranked highly in this component (2,734th out of 19,262), however, it is known to function as a heterodimer with *Msh5*, which ranks 34th (Rakshambikai et al., 2013). This highlights one of the advantages of single cell RNAseq compared to GWAS for target discovery in that it does not rely on the presence of (perhaps rare, small effect) genetic variants. In addition it directly provides a list of genes rather than SNPs affecting unknown causal genes, For example the previous GWAS had identified a SNP in the intron of *Ccdc43*, however our expression data strongly suggested the adjacent gene *Meioc* (aka *C17orf104*) as the causal gene (ranked 183rd vs 13,651st in component 5) in addition to the subsequent reports that *Meioc* is responsible for maintaining an extended meiotic prophase (Abby et al., 2016; Kong et al., 2014; Soh et al., 2017).

The strong enrichment of genes involved in recombination in this component suggests other highly ranked genes of unknown function could also play key roles in this process. This proved to be the case for two such genes during the preparation of the manuscript describing our findings. Firstly *Ankrd31* (ranked 102nd) was found to control the number, timing, and location of double strand breaks in meiosis (Boekhout et al., 2018; Papanikos et al., 2018) and *Hsf2bp* (now *Meilb2*, ranked 194th) was found to be a master regulator of meiotic recombinases (J. Zhang et al., 2019). After the publication of this work, *Hells* (ranked 44th) was found to be

required for PRDM9 binding to hotspots (Spruce et al., 2019). One candidate, *Zcwpw1*, will be the subject of chapter 3.

2.9.9 Pachytene

Pachytene is the longest stage of prophase I (~7 days in mice) and has the highest non-ribosomal RNA synthesis activity (Monesi et al., 1978). Accordingly, a number of components represent this stage and have many highly loading genes.

Many of the proteins expressed at this stage are stored for later use when transcription is repressed due to nuclear condensation. Y-box proteins are abundant in spermatocytes where they bind to and repress translation of mRNA (reviewed in Kleene, 2016). *Ybx2* (aka *Mys2*) has a high loading in component 42 (rank 223), which was found translated from pachytene onwards (Y. K. Kwon et al., 1993; Oko et al., 1996) (transcribed from day 17 & in pachytene cells (W. Gu et al., 1998)) and is required for fertility (J. Yang et al., 2005). *Ybx1* (*Msy1*, rank 5) with reported mRNA detectable by northern blot from day 15 (pachytene) (Tafuri et al., 1993). *Ybx3* (aka *Msy4*, rank 28), also has a high loading and is detected as protein in mid-pachytene cells (H. G. Davies et al., 2000) and extended expression causes infertility (Giorgini et al., 2002)

Poly(A) binding proteins also bind to and regulate mRNA (reviewed in Ozturk et al., 2018). *Pabpc1* and *Pabpc2* both have high loadings in component 42 (rank 4 and 52) and are detectable from pachytene onwards (C. Kleene et al., 1994; W. Gu et al., 1995; Kleene et al., 1998; Y. J. Lee et al., 2000; Kimura, Ishida, et al., 2009). *Pabpc6* also has a high loading (rank 22) but is understudied in comparison.

Spermatocytes have specialist energetic requirements and many core enzymes have different isoforms in spermatocytes. Testis specific lactate dehydrogenase *Ldhc* is ranked highly (1st) in component 42 (Goldberg, 1963; Blanco et al., 1963; Goldberg et al., 2010), in addition to *Ldha* (rank 83) and *Ldhal6b* (rank 19) (H. Wang et al., 2005), and another glycolytic enzyme *Pgam2* (rank 107) for which protein expression was found from pachytene onwards (Fundele et al., 1987).

The earliest pachytene component (#13) has a number of highly loading piwi associated genes: *Piwil1* (Miwi), *Tdrd1* (Tudor), *Tdrd5* (Tejas), *Pld6* (Zucchini), and *Piwil2* (Mili) (rank 4, 7, 8, 42, 84). Other pachytene components also have high loading piwi associated genes such as *Ddx4* (Mvh/Basa) and *Tdrd9* (SpnE) in component #47 (rank 9 and 32), and *Mael* (Maelstrom) in component #39 (rank 42). While the piwi pathway is most commonly associated with transposon silencing, in male germ cells there is a second phase of activity during pachytene where piRNAs originate from piRNA clusters rather than transposable elements, and have been proposed to direct transcriptional silencing of mRNAs (Gou et al., 2014; Iwasaki et al., 2015). Component #39 also has high loadings for many subunits of the chaperonin complex which enables proper folding of actin and tubulin (Dun et al., 2012) (*Tcp1*, *Cct7*, *Cct4*, *Cct3*, *Cct6a*, *Cct5*, rank 23, 117, 124, 212, 325, 333). Most of these genes are most highly expressed in the testis compared to other tissues in GTEx.

Pachytene components can be readily identified by the relative lack of loadings on the X and Y chromosomes (figure 2.16D,E). This is due to a process where homologous chromosomes that fail to synapse are marked by HORMAD1, HORMAD2 and BRCA1. This is followed by recruitment of ATR and the creation of γH2AX resulting in transcriptional silencing (Meiotic Silencing of Unsynapsed Chromatin). As the X and Y chromosomes are only partially homologous they fail to fully synapse and are silenced by this processes (Meiotic Sex Chromosome Inactivation, MSCI) (Turner, 2007; Turner, 2015).

Pseudotime analysis provides quantitative, high-resolution insights into MSCI. We are able to track this process through pseudotime by visualising the ratio of sex to autosomal expression through pseudotime (figure 2.16A & 2.17A). The ratio drops sharply (possibly exaggerated by inefficient cell capture) to almost 0 before gradually recovering, although to a level below that of the original starting ratio. It has previously been reported that some sex chromosome genes are able to escape MSCI (da Cruz et al., 2016; Soumillon et al., 2013). However, we were unable to detect expression of these genes during the time at which MSCI is

active (figure 2.16C & 2.17B). Many of the genes had high expression just before or after MSCI, suggesting that they were previously identified due to the low stage-resolution of the previous study.

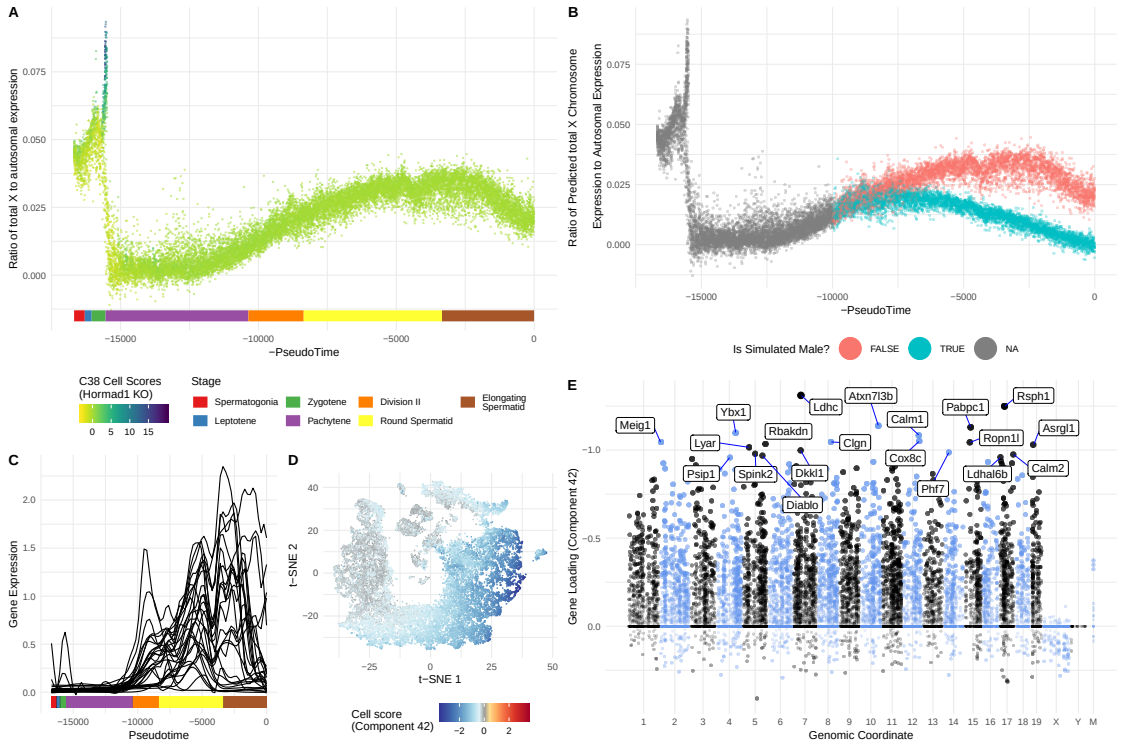


Figure 2.16: **(A)** The sum of imputed expression for all genes on the X chromosome divided by that of the autosomes (y-axis). **(B)** We do not observe that haploid cells obviously split into two populations due to lack of sex chromosome transcript sharing, in part A. Here were simulate what we might expect to see if there was indeed a lack of sharing (methods section 2.12.3). **(C)** Smoothed expression values (unimputed, adaptive GAM smoothing with formula "y ~s(x, bs = ad)") are shown for each gene reported to escape MSCI (da Cruz et al., 2016) excepting *H2al1e*, *H2al1c*, and *Gm10096* which were below our dataset's expression detection threshold. **(D)** Component 42 (Pachytene) cell scores in t-SNE space. **(E)** Component 42 gene loadings. This component represents genes active during the pachytene stage of meiosis; note the striking lack of sex chromosome gene loadings, due to MSCI.

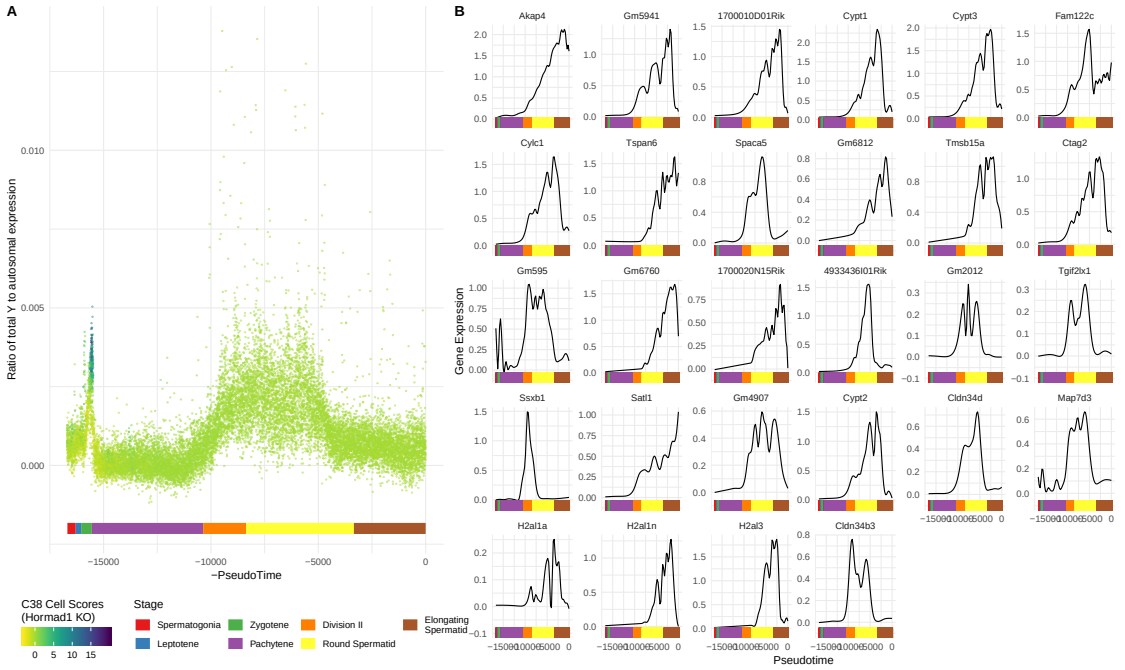


Figure 2.17: (A) As for Figure 2.16A, but Y chromosome instead of X. Due to lower number of expressed genes on the Y chromosome MSCI is not as clear and individual genes can influence the overall ratio to a large degree. **(B)** As for Figure 2.16C, but each gene is shown individually.

In addition to MSCI, we might expect to observe a lack of sex chromosome transcripts due to the lack of either an X or Y chromosome in the haploid cells post meiotic division. However, from the mitotic divisions of the spermatogonia onwards cytokinesis does not fully complete and μm-wide cytoplasmic bridges are formed between adjacent cells such that the cytoplasm could be shared across a syncytium of cells (Greenbaum et al., 2011). Whilst theoretically thousands of cells could be connected in practice up to 650 connected cells have been observed (H. P. Ren et al., 1991). The extent to which mRNA sharing occurs is unknown although efficient sharing has been shown for individual genes (Braun et al., 1989). If many genes were not shared we would expect to see two populations of cells after the meiotic divisions with high and low X chromosome expression respectively. However, we observe a marginal distribution of X expression conditional on pseudotime that is unimodal - suggesting that mRNA (on the X chromosome at least) is efficiently shared (figure 2.16B).

There remains a possibility that some individual or group of genes are not shared, such as has been observed for autosomal genes in a mutant heterozygous context: the t-complex responder mutant (*Smok*^{Tcr}) which functions as an antidote in the poison-antidote meiotic drive system of the t-complex (Herrmann et al., 1999; Véron et al., 2009) and *Spam1* which causes transmission ratio distortion in Robertsonian (Rb) translocation-bearing mice (Martin-DeLeon et al., 2005). Indeed a recent report using hybrid mice claimed widespread haploid biased gene expression (Bhutani et al., 2019).

2.9.10 Hormad1 Sex Chromosome Activation

In addition to the pachytene components with lack of loadings on the sex chromosomes we also find a component (#38) with an *excess* of sex chromosome loadings relative to the autosomes (figure 2.18A). This component is active exclusively in a subset of *Hormad1*$^{-/-}$ cells that diverge from the normal pseudotime around Zygotene stage 2.19C. Note that as *Hormad1*$^{-/-}$ mice arrest at early Pachytene, once the cells diverge the pseudotimes can no longer be compared to the wild type pachytene cells and are perhaps considerably compressed due to the almost perpendicular direction in t-SNE space (Shin et al., 2010). HORMAD1 marks unsynapsed chromosomes, so when HORMAD1 is absent (as in the KO) transcriptional silencing fails to occur. We find that not only does *Hormad1*$^{-/-}$ fail to silence previously expressed sex-linked genes, many previously *unexpressed* sex-linked genes such as *Rhox2h* have high expression (figure 2.18B). Interestingly, there are also multiple autosomal genes with high loadings. This may be due to ectopic expression of sex-linked transcription factors; for example, *Zfy1* and *Zfy2* were previously shown to cause pachytene arrest when forcibly expressed (Royo et al., 2010). We find a very strong association between genes in this component and genes overexpressed in mice which have mutations in either *Hormad1* ($p = 2.2 \times 10^{-39}$, OR = 184) or *Trip13* ($p = 1.3 \times 10^{-157}$, OR = 115) which is required for HORMAD1 removal from synapsed chromosomes (Ortega, 2016; Wojtasz et al., 2009) (figure 2.19A,B).

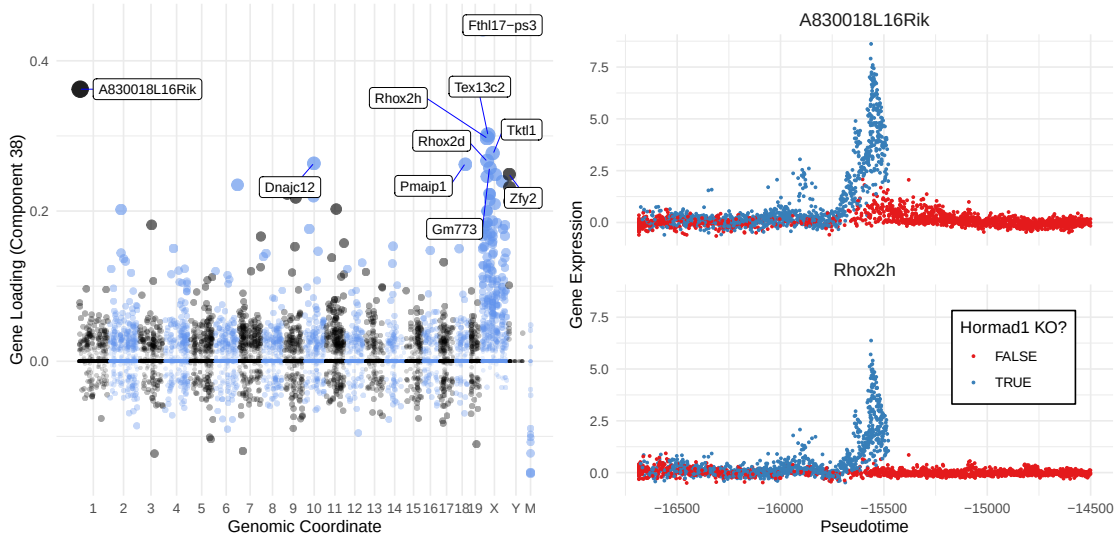


Figure 2.18: **(A)** Gene loadings for component #39 shown along the genome. **(B)** Gene expression for high loadings genes from chromosome 1 (*A830018L16Rik*) and X (*Rhox2h*).

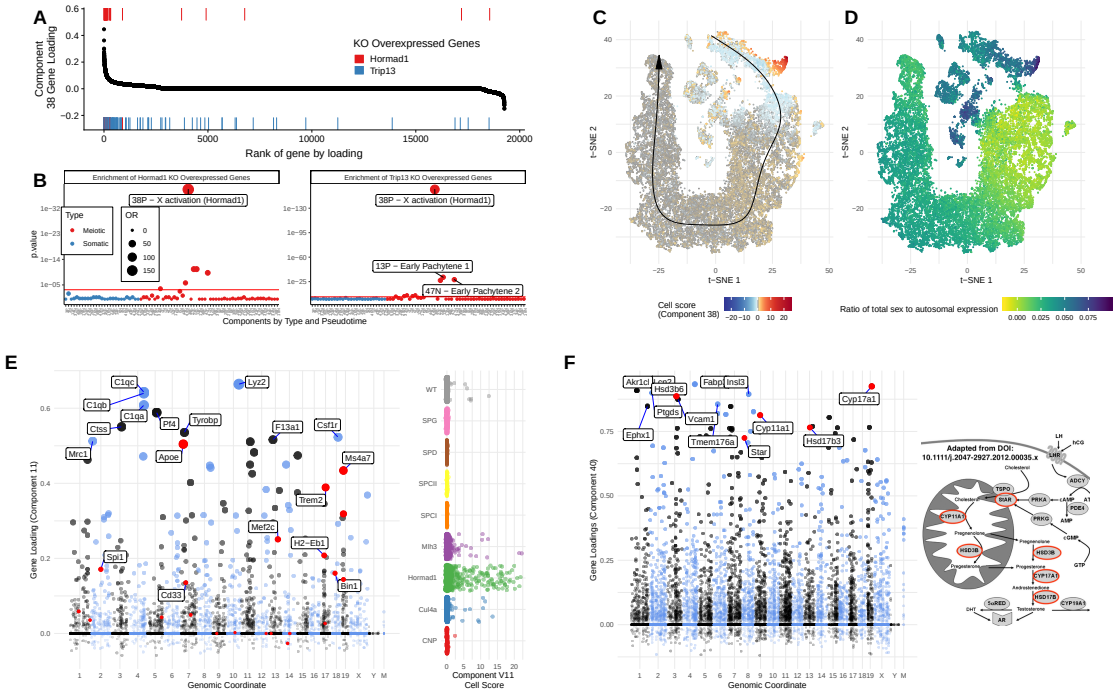


Figure 2.19: **(A)** Enrichment of component 38 gene loadings for genes previously identified as overexpressed in both *Trip13*$^{-/-}$ and *Hormad1*$^{-/-}$ mutants. Component 38 (X Activation) gene loadings are shown ordered by rank (high-to-low) with the genes previously identified as overexpressed in *Hormad1*$^{-/-}$ and *Trip13*$^{-/-}$ mice highlighted as rug plots. **(B)** Fisher's test was used to test for enrichment of genes previously identified as overexpressed in *Hormad1*$^{-/-}$ and *Trip13*$^{-/-}$ mice in the top 500 genes of each component. Positive and negative loadings for each component are analysed separately; thus 38N is the negative gene loadings for component 38 and 38P the positive ones. The positive loadings of component 38 are highly and specifically enriched for these genes. Component enrichments are further coloured red and blue to indicate the primary cell type with loading on each component (red = germ cell, blue = somatic cell). **(C)** Cell scores for component 38 in t-SNE space. The cells with component 38 active are localized to a group of cells which diverge from the pseudotime-line just before pachytene. **(D)** MSCI in t-SNE space. The cells with active component 38 also have a high ratio of sex to autosome expression ratio, showing that not only do these cells fail to silence sex linked genes but they are actively overexpressed. **(E)** Gene loadings and cell scores for component 11 (Macrophages). Genes with GWAS signals for Alzheimer's disease are highlighted in red. **(F)** Gene loadings for component 40 (Leydig cells), with key genes of testosterone synthesis highlighted in red. The pathway of testosterone synthesis is shown with genes that have high loadings in component 40 highlighted (Adapted from Stojkov et al., 2013).

2.9.11 Cul4a, Mlh3, and Cnp Mutants

CUL4A is a major component of the E3 ubiquitin ligase complex called CRL4 which is known to regulate cell cycle, DNA replication, DNA repair, and chromatin

remodelling (Dubiel et al., 2018). Studies on *Cul4a*$^{-/-}$ mice noted that some spermatocytes arrest at the pachytene stage of meiosis I induced by the pachytene checkpoint, whereas remaining spermatocytes complete meiosis but the resulting spermatozoa present oligoasthenospermia and severe malformations (Yin et al., 2011). The molecular basis of the observed abnormal phenotypes in spermatozoa remains unclear. We identified a single SDA component (#25) that was highly specific to *Cul4a*$^{-/-}$ cells. This component corresponds to dozens of genes that are overexpressed in *Cul4a*$^{-/-}$ mutants when compared to all other strains, with GO enrichments related to spermatid development, motility and capacitation. These findings are consistent with the observed phenotype of *Cul4a*$^{-/-}$ mice and provide new leads to investigate mechanisms of pathology.

MLH3 is an essential protein required for crossover formation in early meiosis and for binding of MLH1 to meiotic chromosomes. Studies on *Mlh3*$^{-/-}$ testes have shown depletion of spermatocytes and some spermatogonia due to apoptosis in diplonema induced by a reduction of chiasmata and a loss of recombination nodules (Lipkin et al., 2002). Interestingly, in contrast to *Hormad1*$^{-/-}$, we found no obvious transcriptional phenotype in *Mlh3*$^{-/-}$ cells either by SDA analysis or by comparison of expression levels between hard-clustered wild-type and mutant cells (other than differential expression of *Mlh3*). Instead, *Mlh3*$^{-/-}$ spermatocytes might simply trigger apoptosis through existing checkpoint protein machinery assembled earlier in development (figure 2.20).

Similarly, the cells from *Cnp* mutant mice did not form distinct clusters, nor did they show SDA component loadings distinct from wild-type cells. Although the presence of multinucleated giant cells, hypocellular seminiferous tubules and infertile phenotype in these mice point to a serious defect in spermatogenesis, it is difficult to determine which stages are affected using single-cell expression data. One possible explanation of missing important biological signals may be that *Cnp* mice present a partial arrest phenotype which masks the developmental abnormalities. Another possible explanation is that droplet-based sequencing library preparation

may undersample the cells with aberrant transcriptional signatures, for example due to failure of oil droplets to encapsulate the giant cells.

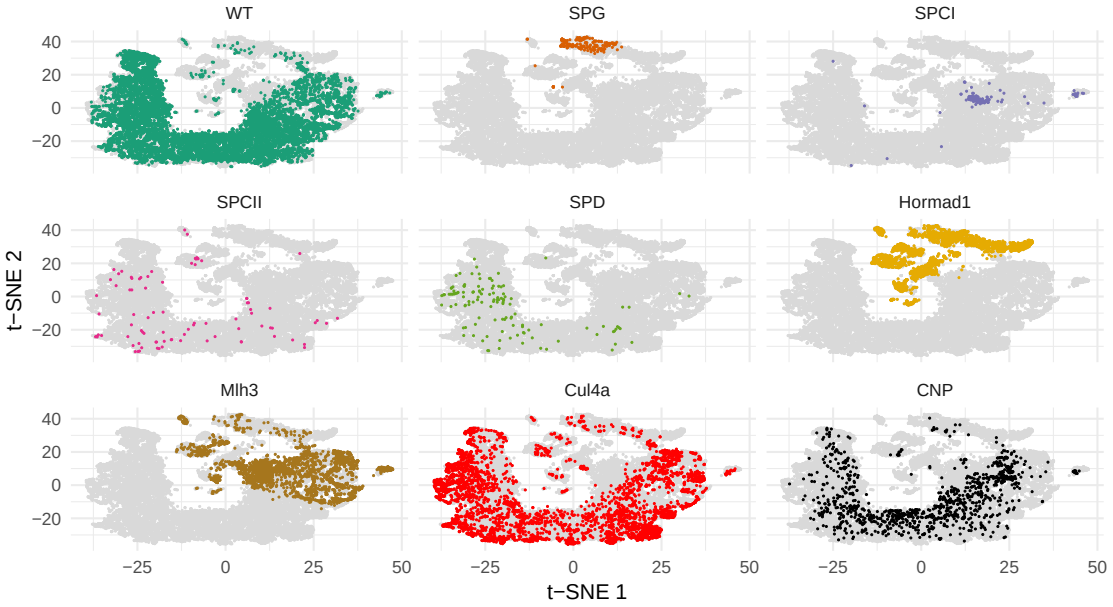


Figure 2.20: Cells are shown projected by t-SNE into two dimensions, with cells coloured by genotype or stage as separated by FACS (SPG: Spermatogonia, SPCI: Primary Spermatocytes, SPCII: Secondary Spermatocytes, SPD: Spermatids).

2.9.12 Meiotic Divisions

Component 20 contains high loadings for cell cycle associated genes such as the cyclin *Ccna1* (rank 45) which is expressed from late pachytene through until just after the meiotic divisions (Sweeney et al., 1996) for which it is required (D. Liu et al., 1998). The cyclin *Ccnb1* controlling the G2/M transition is also present (rank 8), along with its activators *Aurka*, *Bora*, *Plk1*, and *Cdc25a* (ranks 33, 83, 65, and 75) (reviewed in Joukov et al., 2018). The CDK2 inhibitor *Cdkn3* (aka KAP1, (Poon et al., 1995)), *Rgcc* (regulator of cell cycle) which is required for G2/M transition (Saigusa et al., 2007), anaphase promoting complex activator and *Cdc20* homologue *Fzr1* (Holt et al., 2014), and Mitotic Checkpoint Complex antagonist *Mad2l1bp* (*p31comet*) (Habu et al., 2002) also have highly ranked loadings in component 20 (ranks 4, 35, 90, and 118). In addition, the positive and negative loadings

are marginally significantly enriched for "spindle checkpoint" and "chromosome segregation" GO terms respectively (q = 0.029 & 0.015, OR = 10.4 & 3.7).

This component is also highly enriched for a family of proteins with a domain of unknown function (DUF622): *1700001F09Rik*, *Gm3453*, *Gm10354*, *Gm3149*, *Gm8362*, *Gm3127*, *Gm17019*, *Gm4181*, *Speer4e*, *Speer4b*, *Gm9758*, *Gm8232*, *BC061237*, *Gm5458*, *4930572O03Rik*, *Gm5800*, *Gm7361*, *Gm8220* (18 out of the top 88 genes). This family of genes is rodent specific and arose from duplication of the *Dlg5* gene (Church et al., 2009). Interestingly it was reported that genes of this family have similar epigenetic profiles to that of the sex chromosomes in spermatogenesis (Moretti et al., 2016). This component is also host to a second family of ampliconic genes (Ssx - Synovial Sarcoma associated, on the X chromosome), discussed in further detail in section 2.10.3.

2.9.13 Round Spermatid

Round spermatid component 30 contains many genes associated with the acrosome, an organelle which forms a nuclear cap containing hydrolytic enzymes used in fertilization (Ito et al., 2016). This component is highly enriched for "cell wall macromolecule catabolic process" by GO analysis ($q = 1.47 \times 10^{-10}$, OR = 51.6), driven by a family of Chicken-type lysozyme genes: *Spaca1*, *Lyzl6*, *Spaca3* (aka *Lyzl3*), *Lyzl1*, *Lyzl4*, *Spaca7*, *Spaca4*, and *1700016D06Rik* (rank 2, 10, 14, 16, 52, 58, 64, and 92).

SPACA1 (SAMP32) is first detected as protein in mice testis at stage 2 of round spermatids and is required for normal acrosome formation (globozoospermia) (Z. Hao et al., 2002; Fujihara et al., 2012). LYZL6 was detected in primary and round spermatocytes in addition to the post-acrosomal area of spermatozoa (J. Wei et al., 2013).

SPACA3 (*Lyzl3*, *SLLP1*) was detected by gold particle immunostaining to be present at least in the acrosome of human sperm (Mandal et al., 2003), and the anterior acrosome and equatorial segment in mice caudal spermatozoa (Herrero

et al., 2005). Antibodies against SPACA3 prevented sperm egg binding suggesting a role for this protein in fertilisation (Herrero et al., 2005).

LYZL1 was detected in the acrosomal matrix of mouse sperm by mass spectrometry, along with LYZL4, LYZL5, SPACA3 and SPACA5 (Guyonnet et al., 2012). LYZL4 was detected in mouse spermatozoa in the acrosomal region (R. Sun et al., 2011). In situ hybridisation analysis did not detect LYZL4 in human seminiferous tubules but instead in the epithelium of the epididymis (K. Zhang et al., 2005). Lyzl4os (opposite strand antisense lncRNA) ranks 242nd.

SPACA7 expression substantially increases post meiotic divisions as measured by immunofluorescence, and can be detected in the acrosome by immunogold electron microscopy (Korfanty et al., 2012; Nguyen et al., 2014). SPACA4 (SAMP14) showed immunofluorescent staining restricted to the acrosomal region in washed human spermatozoa (Shetty et al., 2003). 1700016D06Rik (aka GC14) contains a Spaca7 domain and was previously reported as expressed in pachytene spermatocytes in contrast with our results (Anway et al., 2003). Other acrosomal genes not covered by the GO term include Lyzl5, Fam170b, Acrv1 (Spaca2), and Spata31 (AEP1) (rank 320, 56, 4, and 30).

Lyzl5 (Spaca5) is an X linked gene not previously studied apart from being detected by mass spectrometry as mentioned above (Guyonnet et al., 2012). Fam170b which showed enrichment in the acrosome of round spermatids by immunohistochemistry and immunofluorescence (Y. Li et al., 2015). Acrv1 (Spaca2) is highly specific to the acrosome as assessed by Immunohistochemistry (Osuru et al., 2014). Spata31 (aka AEP1 - Acrosome Expressed Protein) mRNA was detected by in situ hybridisation from postnatal 25 dpp (round spermatids) in rat (Luk et al., 2006).

Another enriched GO term is "fusion of sperm to egg plasma membrane involved in single fertilization" ($q = 2.1 \times 10^{-6}$, OR = 32.2) driven by *Spata46*, *Lyzl6*, *Spam1*, *Eqtn*, *Hyal5*, *Izumo1* (rank 5, 10, 40, 73, 155 and 201).

Spata46 was detected in elongating spermatids and knock out resulted in abnormal sperm head shape and failure of sperm-egg fusion (J. Chen et al., 2016). *Spam1* (Sperm Adhesion Molecule 1, aka PH-20 aka Hyal1) is a hyaluronidase

originally discovered on guinea pig sperm heads (Myles et al., 1981), and later as required for binding to the egg (Primakoff et al., 1985) and conserved among mammals (Lathrop et al., 1990) but is not required for fertilisation in mice (Baba et al., 2002). Another hyaluronidase, *Hyal5*, is also highly ranked in this component and potentially involved in penetration of mouse sperm (E. Kim et al., 2005; Kimura, E. Kim, et al., 2009). *Eqtn* (Acrosome Formation-Associated Factor), is abundantly expressed in round spermatids and is required for the acrosomal reaction (Y.-C. Li et al., 2006; J. Hao et al., 2014).

Izumo1 and *Izumo2* (rank 201 & 589) are required for sperm to fuse with eggs by binding to the Juno receptor on the egg (Inoue et al., 2005; Bianchi et al., 2014). *Tmem190* (rank 50) was found to co-localise with IZUMO1 (Nishimura et al., 2011) and *Glipr1l1* (rank 86) was found to form a complex with IZUMO1 and is required for optimal fertilisation (Gibbs et al., 2010; Gaikwad et al., 2019).

Zp3r (rank 62) was initially identified as a *Zp3* binding protein (Bleil et al., 1990) and later found to be a component of the acrosomal matrix detectable from late pachytene onwards (K.-S. Kim et al., 2001). However *Zp3* is not required for fertility, like many other proteins involved in sperm-ZP binding (Muro et al., 2012). *Zpbp* (zona pellucida binding protein, rank 241) is an acrosomal protein first detectable in step 2-3 round spermatids and is required for fertility (Y.-N. Lin et al., 2007).

Catsper1, *Catsper3* and *Catsper4* (rank 409, 96, and 162) form a complex (along with *Catsper2*) that functions as a calcium ion channel which is required for sperm hypermotility and fertility (D. Ren et al., 2001; Lobley et al., 2003; H. Qi et al., 2007; J. Jin et al., 2007). *Catsper2* is expressed most highly in Pachytene in our data, ranking 303rd in component 42, with component 30 being the second highest ranked (1995th), consistent with previous reports of premature expression (N. Schultz et al., 2003). One of the accessory subunits *Catsperz* which is required for structural continuity of the Catsper channels along the flagella is also highly expressed in a similar pattern to Catsper2 in our data (Chung et al., 2017).

Adam24, *Adam26a*, *Adam20*, *Adam29*, *Adam25* (rank 126, 133, 262, 278, 454) are intronless genes belonging to a phylogenetic subgroup of ADAM metaloproteases

(I. Choi et al., 2004), and at least one (*Adam24*) is required for normal fertility (G.-Z. Zhu et al., 2009).

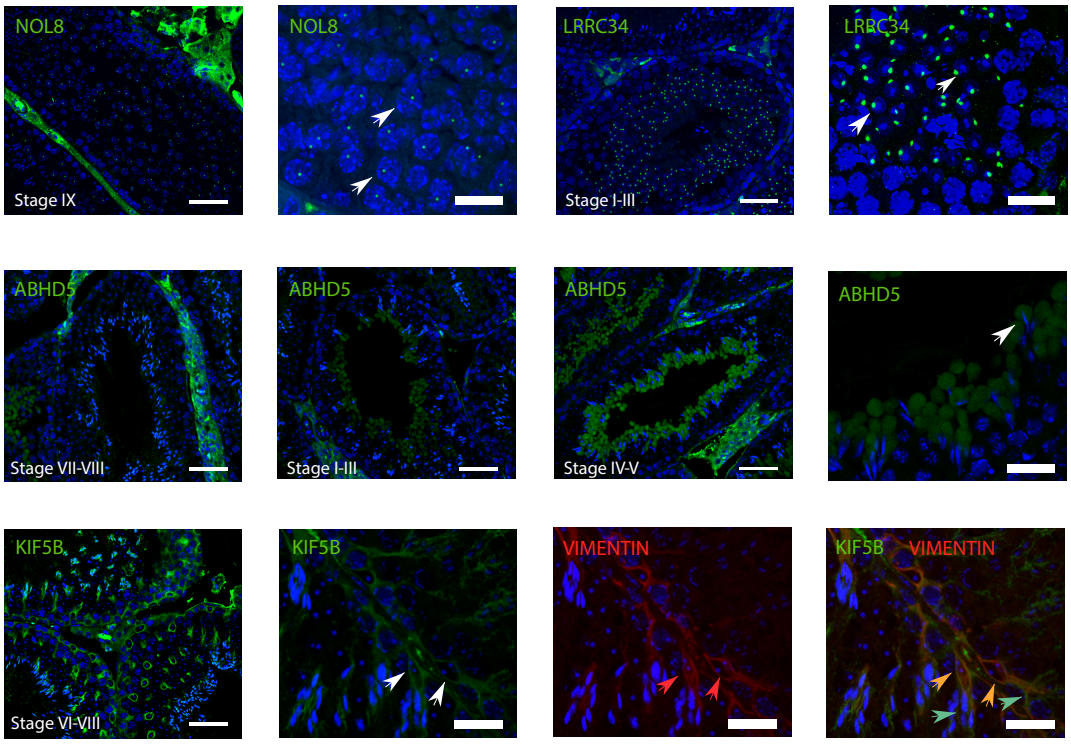


Figure 2.21: This figure is by Min Jung. Thin scale bar, 50 µm; thick scale bar, 20 µm. NOL8, a nucleolar protein, marks primary spermatocytes while LRRC34 marks nucleoli in round spermatids (white arrowheads). Within the tubules, ABHD5 marks specific cytoplasmic regions of elongating spermatids destined to form the residual body (white arrow head) and staining intensity peaks during seminiferous tubule stages IV-V. KIF5B marks Sertoli cells within seminiferous tubules (white arrow head). We co-stained KIF5B with a well-known Sertoli cell marker, Vimentin (red arrow head), and indeed both proteins colocalise to Sertoli cells (orange arrow head). Co-staining also reveals that KIF5B staining extends further out in the cell body (blue-green arrow head) than Vimentin.

Tekt1, *Tekt2*, *Tekt3*, and *Tekt4* are components of flagella (Amos, 2008) and also have high loadings in this component (rank 234, 34, 408, and 101) with expression reported in early pachytene, but with higher expression in round spermatids (Fallahi et al., 2010). *Tssk1* and *Tssk4* (rank 137 & 57) are members of a testis-specific serine kinase family that are almost exclusively expressed post meiotically in the testis (Y. Li et al., 2011). For one high loading gene (rank 145), *Lrrc34*, we verified

by immunofluorescence that the protein is indeed localized to the acrosome of round spermatids (figure 2.21).

Component 35 almost completely overlaps 30 but is strongly multi signed with positive loadings early flipping to negative loadings later in pseudotime. Highly ranked positive loadings in this component include those also very highly ranked (top 100) in the meiotic division component: *Ssxb1*, *Ssxb2*, *Prss46*, *Iqca* and 15 others. This has the result of increasing the expression of these genes during the meiotic division stage which overlaps with the first half of component 35. The opposite is true for the genes with negative loadings in component 35 (although to a lesser extent due to relatively smaller gene loadings compared to component 30) for example *Lyzl6*, *Spaca4*, *Hyal5*, and *Izumo1* have expression biased later than one might expect based on their high loadings in component 30 alone.

Component 15 is active around the (pseudotime) end of component 30 and has high loadings for genes reported to be expressed in round spermatids. *Hemgn*, *Prss37* (required for fertility), and *Calr3* (rank 3, 40, 34) (C. Shen et al., 2013). *Ccdc159* (rank 27) is a coiled-coil domain containing gene that causes infertility with no other significant phenotype in a knock out model and is most highly expressed in the testis in GTEX; however nothing else is known about the function of this gene.

Fam71a, *Fam71b* are both highly and specifically expressed at this stage (ranks 1 and 10). *Fam71b* was previously detected in round and elongating spermatids (Petit et al., 2015) and has been implicated in RNA biogenesis due to binding with FRG1P (van Koningsbruggen et al., 2007). However the exact function of these genes remains unknown.

Interestingly the high loading genes in component 15 are enriched for the GO term "regulation of humoral immune response" (q = 0.01 OR = 17.3) driven by the genes *Cd46*, *Cd55*, *Cd59b*, *Cd37*, *Cd59a* (rank 5, 7, 31, 157, 225) which normally protect autologous tissues from complement mediated lysis. CD46, CD55, and CD59 are present in the inner acrosomal membrane (Cummerson et al., 2006), and disruption of *Cd46* results in accelerated spontaneous acrosome reaction reducing fertility (Inoue et al., 2003). *Cd37* is a member of the tetraspanin family which

includes *Cd9*, well known to be required for sperm-egg fusion (Kaji et al., 2000; Naour et al., 2000; Miyado et al., 2000), however *Cd37* deficient mice are fertile and its role in the testis is unknown (Knobeloch et al., 2000).

There is weak evidence of enrichment for "histone-serine phosphorylation" due to expression of *Smok2b*, *Smok2a*, *4921509C19Rik*, *Gm14151* (rank 44, 221, 112, and 230) which are located together in the genome in pairs and are homologues. In addition *Gm14147* and *Smok4a* (lncRNA) are also highly ranked (312, 453) each located nearby these gene clusters and in the case of *Gm14147* is a homologue of the others at its location.

As with all components, there are many interesting genes with unknown function, some of which have previously characterised expression profiles including *Tex33*, *Fam209*, and *Ccer1* (rank 73, 239, 196) (J. T. Kwon et al., 2017).

2.9.14 Spermiogenesis

The spermiogenesis components 17, 18 and 34 all contain many genes known to be expressed at the latest stages of spermatogenesis. Component 17 is broad and positive, whereas 18 splits 17 into two, and 34 has positive loadings in the middle of component 17. The protamines and transition proteins which replace the histones are the most highly expressed genes at this stage: *Prm1*, *Prm2*, *Prm3*, *Tnp1*, *Tnp2* (Sassone-Corsi, 2002), with rank 1st, 50th, 2nd, 3rd, and 4th in component 17 respectively, in addition to high loadings in the other components e.g. 17th for *Prm3* in component 34. Component 17 is highly enriched for "spermatid development" and "flagellated sperm motility" ($q = 4.0 \times 10^{-7}$ and 2.1×10^{-6}, OR = 7.1 and 9.1).

H1fnt (aka *H1t2*, rank 3, 22, and 27 in component 36, 17, and 34) is a variant of the linker histone H1 which is specifically expressed in haploid spermatocytes and is required for DNA condensation and hence normal fertility (Martianov et al., 2005; Tanaka et al., 2005). *Hils1* is another H1 variant exclusively expressed in elongating and condensing spermatids and reported to be required for fertility specifically on a *Tnp2* knockout background (rank 2, 20, 110 in component 18, 34, 17) (Yan et al., 2003; Q. Wu et al., 2009).

Spermatocytes have specialised energetic requirements and have developed correspondingly specialised apparatus to support this. While mitochondria, the site of oxidative phosphorylation, are localised to the midpiece glycolytic enzymes are localised to the fibrous sheath along the principal piece (Westhoff et al., 1997; Krisfalusi et al., 2006). This suggests glycolysis is required to satisfy the ATP demands far away from the mitochondria and indeed chemical inhibition of glycolysis resulted in reduced motility whereas mitochondrial inhibition did not (Mukai et al., 2004). Many of the enzymes involved in the central metabolic pathway of glycolysis have altered expression or specific isozymes during spermatogenesis and especially spermiogenesis. The first step in glycolysis is catalysed by Hexokinase encoded by *Hk1* (rank 29, 51, 229 in components 34, 18, and 17), which is expressed as a testis specific isozyme lacking the porin domain in spermiogenesis (Mori et al., 1993; Nakamura et al., 2008). The second step is catalysed by Glucose-6-phosphate isomerase encoded by *Gpi1* (rank 173, 200, and 283 in components 18, 34, and 17) which also has a splice variant specifically expressed in spermatogenesis (Buehr et al., 1981; Vemuganti et al., 2010). *Aldoart2* is a rodent specific retro-transposed version of *Aldoa* encoding aldolase which catalyses the fourth step of glycolysis (rank 241 and 264 in components 18 and 17) (Vemuganti et al., 2007; Vemuganti et al., 2010). The products of this reaction can be interconverted by Triosephosphate isomerase which is encoded by *Tpi1* (with ranks 250, 339 and 469 in components 18, 34 and 17) and also has testis specific isozymes (Ijiri et al., 2013). The sixth and seventh reactions are catalysed by GAPDH and PGK respectively, and there are testis specific versions of the genes encoding these enzymes (*Gapdhs*, *Pgk2*) that are required for motility and fertility and have high rankings in these components (rank 3 and 46 for component 18, rank 10 and 73 for component 17, and rank 146 and 161 for component 37) (McCarrey et al., 1987; J. E. Welch et al., 1992; Miki et al., 2004; Danshina et al., 2010). *Pgam2* encoding PGAM-B which catalyses the 8th step, is specifically expressed from pachytene onwards but not earlier (rank 183 and 283 in component 17 and 18) (Fundele et al., 1987). Lactate is produced from the final metabolite of glycolysis by lactate dehydrogenase. There is a testis

specific gene *Ldhc* which is required for fertility, WT motility, and hyperactivation, and whilst its expression starts earlier than the other isozymes is still expressed in the later stages with gene loading ranks of 149, 188, 197 in components 34, 17, and 18 (Sakai et al., 1987; Odet et al., 2008).

In addition, *Abhd5* (aka CGI-58), a protein previously detected in testis lipid droplets (W. Wang et al., 2015), has high loadings specifically in these late components (rank 98 and 313 in components 18 and 17) and we show by immunofluorescence that it serves as an excellent marker of the residual body (figure 2.21).

2.10 Motif Inference

We hypothesized that many of the SDA components represent dynamic and finely tuned transcriptional programs. One major class of transcriptional regulation is the binding of transcription factors to promoters and enhancers. If it's true that transcription factors are driving the co-expression represented by the components, we would expect that the targets of known transcription factors would be enriched in specific components contiguous in developmental time.

2.10.1 Enrichment of Known Transcription Factor Targets

There are some transcription factors that are known to regulate meiotic gene expression. One of these is MYBA (encoded by *Mybl1*, Bolcun-Filas et al., 2011) and we find that the predicted direct targets of the MYBA transcription factor (as found by ChIPseq and RNAseq of KO mice) are strongly enriched specifically in the pachytene components as expected based on known targets and expression of *Mybl1* (figure 2.22). Other known meiotic transcription factors including *Stra8* (Kojima et al., 2019), *Rfx2* (Kistler et al., 2015), and *Crem* (Nantel et al., 1996) are also enriched in specific components contiguous in pseudotime and consistent with their known targets.

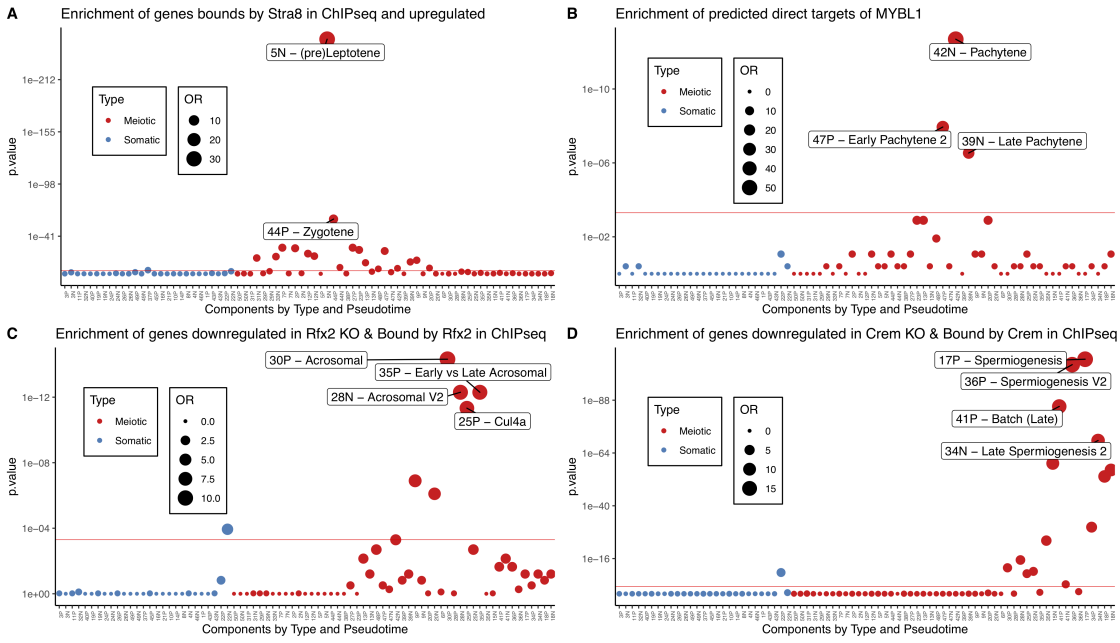


Figure 2.22: As part of the validation that some SDA components represent biological co-expression we tested for enrichment of ChIP-seq defined target genes (methods section 2.12.5) of well-known meiotic transcription factors (A) STRA8, (B) MYBL1, (C) RFX2, and (D) CREM. Enrichment was calculated using Fisher's test against the top 500 genes in each component (positive and negative loadings separately). OR = Odds Ratio. Red line represents Bonferroni corrected p = 0.05.

2.10.2 *De-novo* inference from promoter sequences

Transcription factors are able to recognise particular subsets of gene promoters by sequence specific binding to DNA motifs. Hence, if transcription factors are responsible for the co-expression patterns observed, we would expect to see enrichment of these binding motifs in the promoters of the co-expressed genes. However, it is possible that there are undiscovered transcription factors and or binding motifs. Therefore instead of looking for enrichment of known motifs we inferred motifs *de novo* from the promoter sequences of the top genes from each component using an existing algorithm (Altemose et al., 2017; B. Davies et al., 2016).

This resulted in 16 groups of motifs (figure 2.23 and methods). By comparing these to known motifs in the HOCOMOCO database, we found matches to the master regulators *Stra8* (only recently discovered), *Mybl1*, *Rfx2*, *Crem*, as well as other genes required for fertility *Nrf1*, *Etv5*, *Zfp143*, and *Rara*. In addition, we identified the target of *Spi1* (aka PU.1), a known master regulator of macrophage

differentiation (Rosa et al., 2007), specifically in the macrophage component 11. High and specific enrichment of ChIP-seq targets of STRA8, MYBL1, RFX2 and CREM validated our interpretation - that covariation of expression of genes within many components reflects shared transcriptional regulation.

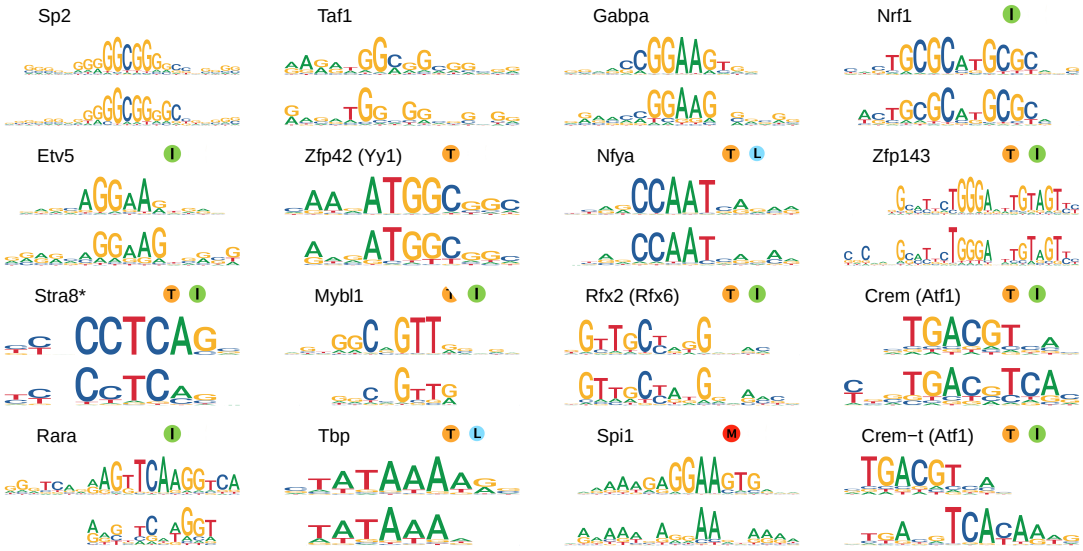


Figure 2.23: Motifs discovered from the promoter sequences of genes with high component loadings. In each motif logo pair the lower logo shows the de novo inferred motif and the upper logo shows the motif in the HOCOMOCO database best matching the de novo motif. Orange 'T' indicates this transcription factor is highest expressed in testis in the GTEx database (half T indicates second highest). Green 'I' indicates that a mouse knockout of this gene is infertile. Blue 'L' indicates a mouse knockout of this gene is embryonic lethal. Red 'M' indicates this gene is required for macrophage development. The notation 'Crem-t (Atf1)' indicates that we suspect that the true transcription factor recognizing the motif is not the closest matching database-motif (Atf1). * the (upper) STRA8 motif shown is from Kojima et al., 2019, rather than the HOCOMOCO database.

2.10.3 Broad-scale patterns of TFBS occurrence

Many motifs were found from multiple components and so to gain a broader view of which components these motifs act in we computed the linear association between the gene loadings and the motif presence probability at the promoters of the genes (figure 2.24). The known master regulators showed relatively specific enrichment (e.g. *Stra8* in Leptotene component #5, *Mybl1* in Pachytene component #42, and *Rfx2* in Acrosomal component #30). Most motifs, however, have a broad pattern of enrichment in either the diploid components or the haploid components. Other

mechanisms which could control the fine-scale patterns of transcription we observe are transcription factor binding at enhancers (in which we did not look for motifs given their relatively unknown locations and gene associations), chromatin status (modulating transcription factor binding), and or differential mRNA degradation.

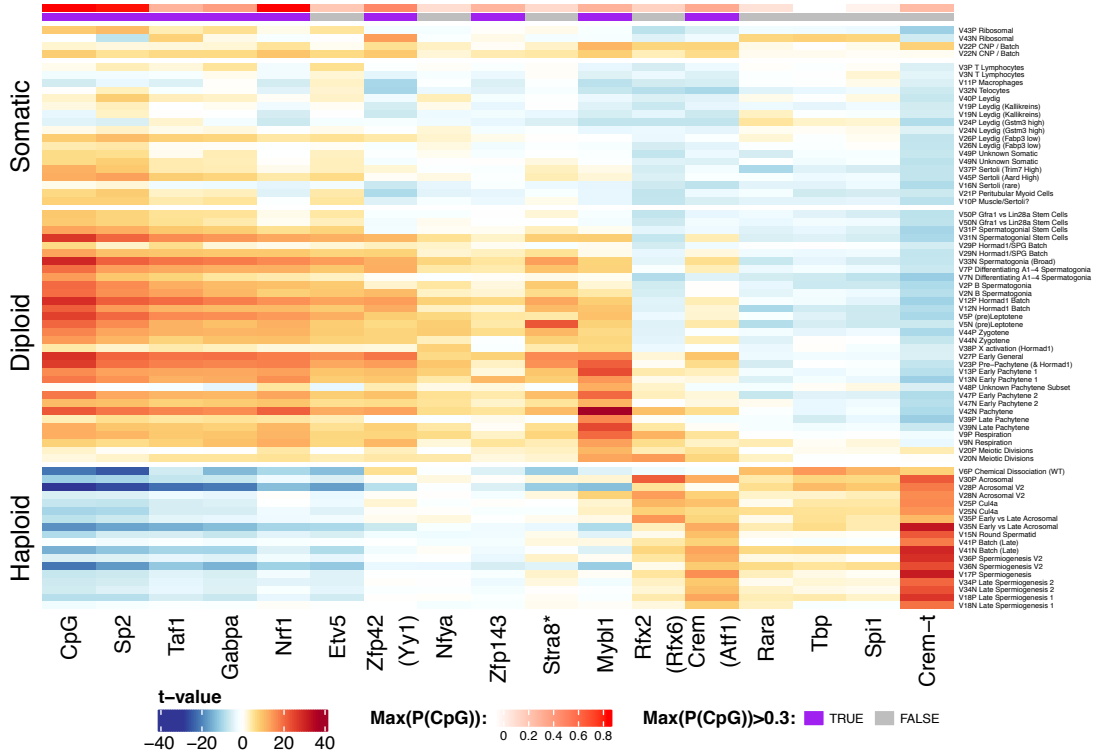


Figure 2.24: Association of gene loadings with the probability each de novo identified motif is found in the genes for each component. Colouring is a Z-score from a correlation test between gene loadings and motif probabilities, where red (blue) indicates positive (negative) association. The germ cell components (rows) are ordered by pseudotime. The correlation was calculated for positive and negative parts of the component separately and in the cases where the component is mainly one-sided the other side has been omitted, as have the single cell components. The additional column 'CpG' shows the same association test, but with count of promoter CpG dinucleotides, for each component. Across the top of the panel, colour bars indicate the maximum probability of there being a CpG at any one position in the denovo motif, and whether that probability is greater than 0.3.

We discovered that many of the motifs which show similar associations through pseudotime share a common property: the presence of a CpG dinucleotide. Many of the motifs with association overlapping diploid components contain a CpG dinucleotide whereas the motifs mostly overlapping haploid components do not. In fact the motif with the strongest association with pseudotime is not a motif

per-se but simply the count of CpGs at the promoter, indicating a major shift away from expression of genes whose promoters contain CpG islands, following the meiotic divisions. CpG islands at active promoters are usually unmethylated and methylation can cause silencing (E. Li et al., 2014). Many of the diploid CpG containing motifs have methylation sensitive transcription factors, including the non-CpG exceptions NFYA and ETV5 (Domcke et al., 2015; J. Wang et al., 2017). This suggests that methylation may play a role in the mechanism by which the type of transcription factor used pre and post division undergoes this switch.

There are many genes that are expressed around the time of the meiotic divisions in component 20 and so could be effectors of this switch. However, one family of genes (*Ssxb1*, *Ssxb2*, *Ssxb3*, *Ssxb5*, and *Ssxb6*) are all highly ranked (top 150) in this component. These genes are known to contain both an SSXRD and KRAB-related domain, the same two domains which are also present in *Prdm9* (the only such gene for which this is true outside the Ssx X chromosome cluster). The KRAB-related domain has been studied in the context of *Prdm9* where it was found to interact with a number of other proteins including the CpG binding protein CXXC1 (Imai et al., 2017; Parvanov et al., 2017). The SSXRD domain has been studied in the context of the Ssxb family's involvement in synovial sarcomas, where it was found to interact with the CpG binding protein CXXC2 (aka KDM2B) (Banito et al., 2018). Both CXXC1 and CXXC2, as their name suggests, are members of a protein family containing the ZF-CxxC domain which mediates recruitment to unmethylated CpG islands (reviewed in Long et al., 2013). CXXC1 is a component of the SETD1 methyltransferase complex that catalyses H3K4me3 at CGI promoters (J.-H. Lee et al., 2005). CXXC2 is a demethylase which actively removes H3K36me2 at CGI promoters in addition to recruiting polycomb repressive complex 1 which in turn catalyses H2AK119ub1 (J. He et al., 2008; Farcas et al., 2012; J. He et al., 2013; X. Wu et al., 2013). The repressive histone modification H3k27me3 has been reported to increase substantially from pachytene to the round spermatid stage, and this mark can be deposited by another polycomb family complex PRC2 (Sin et al., 2015).

2.10.4 Novel Motifs

Some motifs have relatively poor matches to known motifs in the HOCOMOCO database. For example one motif is most similar to the ATF1/CREM motif, but lacks the central CpG dinucleotide and has an additional CAA tail (figure 2.23). It is known that CREM is active as an alternative isoform (tau) in late spermatogenesis, and so it is possible that the alternative motif we discovered is the true motif bound in vivo (Sassone-Corsi, 2000). The associations with components are much more cleanly separated into pre and post division for the new motif, being much stronger post division, consistent with the lack of CpG compared to the database motif (figure 2.24).

2.11 Interactive Web App

In order to make these results widely available to others in the spermatogenesis community, we created an interactive web app which enables users to browse the pattern of gene expression for any gene, as well as explore which genes are in which components, without having to download or install any software themselves (figure 2.25). The app is based on the RShiny framework and is hosted using Amazon Web Services EC2 t2.micro instance (1CPU, 1GB RAM). The website can be viewed at `www.testisAtlas.ml` in addition to being archived within the github repository.

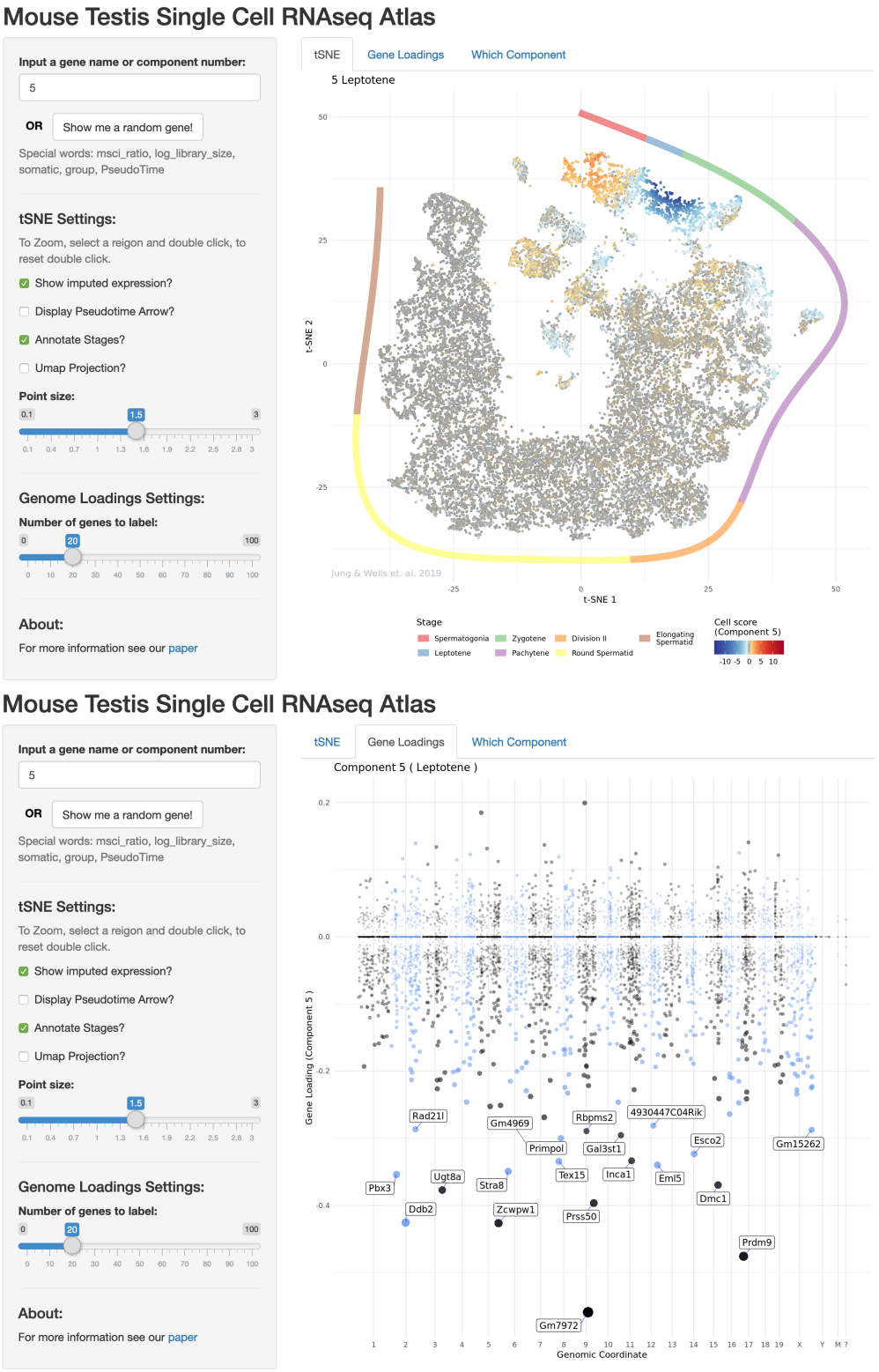


Figure 2.25: Screenshots of the testisAtlas.ml website exemplifying two of the main functionalities of a) viewing component scores/gene expression in t-SNE space and b) exploring gene loadings of components.

2.12 Detailed Methods

For experimental methods please refer to Jung et al., 2019.

2.12.1 Preprocessing of Drop-seq data

Paired-end sequencing reads were processed, filtered and aligned as described in Macosko et al., 2015. The specific steps and tools for this process is further outlined in Drop-seq Computational Cookbook (`http://mccarrolllab.com/wp-content/uploads/2016/03/Drop-seqAlignmentCookbookv1.2Jan2016.pdf`). STAR aligner was used to map the processed reads to mouse genome (Dobin et al., 2013). A STAR indexed genome was generated using the mm10 mouse genome and GRCm38 gene annotation (release version 76) with default settings. Following the alignment, digital gene expression (DGE) matrices were generated for each experimental batch (Macosko et al., 2015).

2.12.2 Software

Custom R code used is available at `www.github.com/MyersGroup/testisAtlas` and archived at DOI: 10.5281/zenodo.3233958. SDA is available from `https://jmarchini.org/sda/`.

Computational analysis was performed using R (R Core Team, 2018). Gene ontology enrichment analysis was performed on the top 250 genes from each component (from each side) using the enrichGO function from the clusterProfiler R package in which p-values are calculated based on the hypergeometric distribution and corrected for testing of multiple biological process GO terms using the Benjamini-Hochberg procedure (Yu et al., 2012). Plots were created using the ggplot2 package and extensions ggrepel, ggforce, ggseqlogo, ggnewscale, ggrastr, RColorBrewer, viridis, and cowplot (Campitelli, 2019; Garnier, 2018; Neuwirth, 2014; Pedersen, 2016; Petukhov, 2018; Wagih, 2017; Wickham, 2016; Wilke, 2018). In addition the following R packages were used: data.table, Matrix (for sparse large matrix computations), biomaRt (for gene identifiers), shiny and shinycssloaders (for creating

the interactive web application), ComplexHeatmap, bigmemory (for creating a low-memory shiny app), and MASS (for kernel density estimation) (Bates et al., 2018; Chang et al., 2018; Dowle et al., 2019; Durinck et al., 2005; Z. Gu et al., 2016; Kane et al., 2013; Sali, 2017; Venables et al., 2002). Procrustes rotation was performed using the R package vegan (X. Lin et al., 2019; Oksanen et al., 2019).

2.12.3 X Chromosome Non Sharing Simulation

To calculate simulated haploid non-sharing (figure 2.16B) cells with pseudotime > -10000 were randomly split into two groups. The predicted X expression was calculated as Original X $\times$ PseudoTime / 10000 e, where e is a random normal error with mean 0 and s.d. of 3.

2.12.4 Imputation Validation

Imputed gene expression values (the posterior means of the SDA model) were computed as the matrix product of the cell scores and gene loadings matrix from SDA.

In order to formally quantify the accuracy of SDA imputation, we performed a cross validation study comparing the ability of SDA imputation to correctly predict single cell gene expression data in a withheld sample. First, we randomly split the post-QC RNA-sequencing reads from the full dataset into two batches: with 20% probability a read is assigned to the test dataset, and with 80% probability it is assigned to the training dataset. Next we create seven predictors of gene expression levels for each cell, using the training dataset: 'Unimputed' uses the training data directly (scaled by the total UMI counts for each cell), 'Mean cell' uses the sum of training reads for each gene across all cells to predict ranks (i.e. every cell has the same prediction), the matrix factorization approaches SDA, ICA, PCA and NNMF were run on the normalized training data (normalized as described above) and imputed values calculated as the matrix product of cell scores and gene loadings, MAGIC values were computed using the Rmagic package (van Dijk et al., 2018).

To compare the accuracy of the three predictors for gene expression imputation, we evaluate an objective function for each predictor and each cell, which we call the

'rank prediction accuracy curve' or RPAC. The RPAC for each predictor is created by rank ordering all genes in a single cell by the predicted level of expression of those genes, from high-to-low, after reversing normalisations (figure 2.9). For each rank (abscissa), the ordinate is the cumulative fraction of test data reads for all genes up to that rank (i.e. all genes with higher predicted expression than the current rank). The RPAC is similar in spirit to a receiver operating characteristic (ROC) curve. The area under the curve (AUC) for each RPAC is informative about prediction accuracy; a completely random predictor is expected to produce an AUC of 0.5, while a method with some predictive utility will have an AUC >0.5. This allows us to prefer predictions with a higher AUC, although we note that (unlike for a ROC) even given perfect imputation, the maximum possible expected AUC is <1, because the test data is sparse and so shows considerable noise relative to the unknown truth.

In order to identify differences between SDA and NNMF (the most similar alternative method), for each gene we calculated imputed expression for both methods (not using single cell components from SDA), and calculated Pearson correlation between the two methods. We then looked for enrichment (by FET, p = 0.05 after correction for multiple testing by Bonferroni) of the 500 least correlated genes in both SDA and NNMF components, finding seven enriched SDA components (3P, 16N, 4N, 10P, 50N, 46N, 8P) and 0 NNMF components. We show example genes from 3P and 50N in figure 2.9D and F.

NNMF analysis was performed using the NNLM R package (X. Lin et al., 2019) with 50 components, and a stop criterion of 10^{-5}. ICA analysis was performed with the fastICA R package (Marchini et al., 2017) with 50 components. PCA was performed using the R package flashpcaR with 50 components and divisor and standardization set to 'none' (Abraham et al., 2014). MAGIC was performed with default parameters using the R package Rmagic 1.5.0. (van Dijk et al., 2018).

2.12.5 Transcription Factor Motif Analysis

To discover de novo motifs enriched within each component we used the MotifFinder software - an iterative Gibbs sampler described in Altemose et al., 2017; B. Davies

et al., 2016 and available at `www.github.com/myersgroup/MotifFinder`. We ran the MotifFinder R package on the repeat masked promoter sequences from Mus Musculus GRCm38 of the top 250 positive and negative genes (separately) for each component. Sequences with greater than 10% masked bases were removed. For each component 10 different regions around the TSS were used (150, 200, 250 bp upstream and downstream of the TSS (separately), and 200, 300, 400, and 800 bp centred on the TSS). Each run was repeated nine times with different random seed motifs (each of length 6 bp). In each case MotifFinder was run for 1000 iterations or until convergence (defined as when standard deviation of the motif proportion is below 0.05).

The resulting de novo motifs were annotated with known motifs from the HOCOMOCO database (V11) using Tomtom from the MEME suite (Bailey et al., 2015; S. Gupta et al., 2007; Kulakovskiy et al., 2018). We subset the matches by taking the match with the minimum q-value for each HOCOMOCO target for those with a E-value of less than 0.001. This resulted in 123 different matches from 101 de novo motifs. These were manually grouped into 16 categories based on motif similarity. Within each group a single 'most likely acting' motif was chosen based on external suggestive evidence, including whether a knockout of the transcription factor (TF) causes infertility, if the TF is specifically expressed in testis by RNA-Seq data in GTEx (GTEx Analysis Release V7 (dbGaP Accession phs000424.v7.p2)), or if the TF is previously known to bind testis expressed genes by ChIP-Seq (GTEx Consortium, 2013).

To find good motifs with poor matches to currently known motifs we plotted the sum of the information content by the E value for each de novo motif. This identified a set of similar motifs with poor E values but large total information content – most of these matched best to ATF1.

In order to determine patterns of association of motifs with pseudotime, we performed correlation tests (using R's cor.test) between the gene loadings of each component (positive and negative loadings separately) and the probability of the motif being present at the promoter of these genes. To find probability of

motif presence per gene promoter, we used MotifFinder but fixed the position weight matrix to either one of our de novo motifs (figure 2.23) or a motif from the HOCOMOCO database and ran MotifFinder for 20 iterations. We assessed convergence through the change in proportion of sequences containing the motif.

We used published ChIP-Seq data for the transcription factors *Stra8* (Kojima et al., 2019), *Mybl1* (Bolcun-Filas et al., 2011), *Crem* (Kosir et al., 2012), and *Rfx2* (Kistler et al., 2015) to validate our conclusion that some SDA components correspond to genes coregulated by these transcription factors. We performed a Fisher's exact test on the overlap of the genes suggested from the ChIP-Seq studies for each of the three transcription factors with the top 500 genes for each SDA component (positive and negative loadings separately). For *Stra8* the genes are those described as 'STRA8-activated' from supplementary table 1 of Kojima et al., 2019. For *Mybl1* the genes are those determined as potential direct targets of MYBL1 (those that were bound in ChIP and were mis-regulated in repro9 mice, from Table 1 of Bolcun-Filas et al., 2011). For *Crem* the genes were those that are "found in the cross-section between DE genes from Kosir et al., 2012 and genes bound by CREM in testis from Martianov et al., 2005 (i.e. the top 50 genes from Kosir et al., 2012 Supp Table 4). For *Rfx2* the genes are the genes from Kistler et al., 2015 that are bound by RFX2 by ChIP-seq and are statistically significantly downregulated at P30 (from their Table S1).

2.12.6 Data availability

Raw data and processed files for Drop-seq experiments are available under GEO accession number GSE113293. Processed data is available at DOI: 10.5281/zenodo.3233870.

The association of paternal and maternal chromosomes in pairs and their subsequent separation during the reducing division as indicated above may constitute the physical basis of the Mendelian law of heredity.

— Sutton, 1902

3 Zcwpw1

Contents

3.1 Introduction 95
3.2 ZCWPW1 co-evolves with PRDM9 97
3.3 ZCWPW1 localisation and KO phenotypes 99
3.4 ZCWPW1 recruitment by PRDM9 104
3.5 DSBs positioning and repair 109
3.6 ZCWPW1 binds CpG dinucleotides 113
3.7 Discussion 116
3.8 Methods 120
3.8.1 Orthologue Alignment 120
3.8.2 Domain Search 122
3.8.3 Mice and genotyping 123
3.8.4 Fertility measurements 124
3.8.5 Immunostaining of spermatocytes 124
3.8.6 FISH 125
3.8.7 ZCWPW1 antibody production 125
3.8.8 Cell line, transfection, and IF staining 126
3.8.9 Immunoprecipitation and Western Blot detection 126
3.8.10 ChIP-seq 127
3.8.11 Data and Code availability 132

3.1 Introduction

Using single cell RNA-sequencing of mouse testis, we identified a set of genes co-expressed in (pre)leptotene cells which are highly enriched for genes involved in

meiotic recombination (section 2.9.8). *Zcwpw1*, which ranks 3rd in this set after *Prdm9* (2nd), is of unknown function but contains two recognised protein domains: CW and PWWP, shown to individually bind H3K4me3 and H3K36me3 respectively (F. He, Muto, et al., 2010; Rona et al., 2016). PRDM9 deposits both H3K4me3 and H3K36me3 at sites it binds (Powers et al., 2016), and this methyltransferase activity is essential for its role in double strand break positioning (Diagouraga et al., 2018), suggesting these marks may be recognized by downstream protein(s). We therefore hypothesised that ZCWPW1 might co-localise to recombination hotspots, by binding the histone modifications deposited by PRDM9.

In humans, ZCWPW1 is specifically expressed in testis (Carithers et al., 2015; Uhlén et al., 2015) (figure A.1). It is also one of 104 genes specific to meiotic prophase in murine fetal ovary (Soh et al., 2015), further suggesting a conserved meiotic function. This was confirmed by a recent study showing that ZCWPW1 is required for male fertility, with ZCWPW1 hypothesised to recruit the DSB machinery to hotspot sites (M. Li et al., 2019). Here, we show that ZCWPW1 co-evolves with PRDM9, and is recruited to recombination hotspots by the combination of histone marks deposited by PRDM9. However ZCWPW1 is not required for the positioning of DSBs at PRDM9-bound sites, which occurs normally in ZCWPW1-null mice. Instead, ZCWPW1 is required for proper inter-homologue interactions: synapsis and the repair of DSBs. In ZCWPW1-null mice, DMC1 signals show strong perturbations, with signals at autosomal hotspots resembling those on the X-chromosome, which does not have a homologue. Thus, ZCWPW1 represents the first protein directly positioned by PRDM9 binding, but impacting homologous DSB repair.

All experimental work in this chapter was performed by Emmanuelle Bitoun, Daniela Moralli, and Gang Zhang. Anjali Hinch helped with SSDS processing, and section 3.2 is by Simon Myers (apart from figure 3.1). Much of this chapter has been published Wells et al., 2020, on which I am co-first author.

3.2 ZCWPW1 co-evolves with PRDM9

Based on publicly available databases, we identified likely ZCWPW1 orthologues in 167 species in total, aligned each to the human reference ZCWPW1 protein, and compared against a previous analysis of PRDM9 (method section 3.2, Z. Baker et al., 2017). The alignment reveals the regions containing the CW and PWWP domains to be the most conserved among species (Figure 3.1), while other parts of the protein appear to have been lost in some species. In addition, there is a region of moderate conservation downstream of the PWWP domain, not overlapping any known domain. Notably, an SYCP1 (SCP1) domain is annotated in the mouse protein only, which although only suggestive, is interesting given that SYCP1 physically connects homologous chromosomes in meiosis. In addition, protein threading suggests that the C terminal end of ZCWPW1 may contain a methyl-CpG binding domain (methods section 3.8.2, Lobley et al., 2009).

Because of incomplete available DNA and protein sequences, we are almost certain to miss some species where ZCWPW1 is present, and similarly it is challenging to identify PRDM9 orthologues (Z. Baker et al., 2017). Despite this, we see extremely high overlap between PRDM9 and ZCWPW1 occurrence. All identified ZCWPW1 orthologues are in vertebrates, similar to identified PRDM9 orthologues (Z. Baker et al., 2017). 99 of 167 (149) species with ZCWPW1 (PRDM9) respectively have the other protein; an even higher fraction of species have a close relative with an identified orthologue: 131 species with ZCWPW1 are in a family possessing PRDM9. Only 8 species possess ZCWPW1 but are in an order with no identified species possessing PRDM9: three amphibians including Xenopus frogs, and five placental mammals. Given widespread conservation of PRDM9 among mammals other than canids (Z. Baker et al., 2017), these five might plausibly represent false negatives. Thus, ZCWPW1 appears to mainly occur in species, and even more clearly in groups of species, possessing PRDM9.

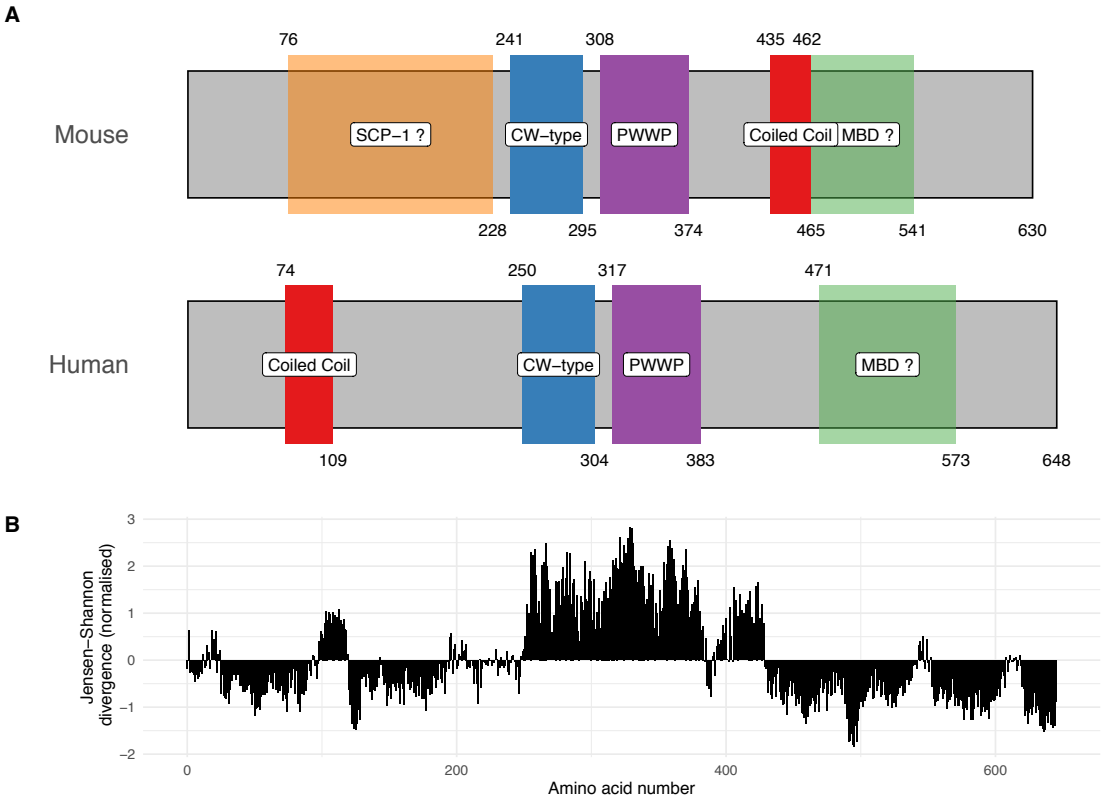


Figure 3.1: Domain organisation (A) and evolutionary conservation (B) of ZCWPW1. **(A)** Protein domains in the human and mouse proteins (source: UniProt). Start and end positions of each domain are shown above and below the rectangles respectively. Prediction of SCP-1 domain from (Marchler-Bauer et al., 2004) and of MBDs from (Lobley et al., 2009) (methods section 3.8.2) **(B)** Conservation of human amino acids, normalised Jensen-Shanon divergence from (Capra et al., 2007; Johansson et al., 2010) using multiple alignment of 167 orthologues (methods section 3.2).

We identified ZCWPW1 orthologues across the full spectrum of vertebrates possessing PRDM9, including jawless, cartilaginous and bony fish, coelacanths, turtles, snakes and lizards, and mammals. Previous work (Z. Baker et al., 2017) has identified at least 6 independent complete PRDM9 loss events; in birds and crocodiles; in three distinct groups of teleost fish (although these fish possess a PRDM9 "beta" orthologue with mutations in key catalytic amino acids within the SET domain); in canids; and tentatively, in amphibians. The first four losses appear to have been completely mirrored by corresponding ZCWPW1 losses, with no species previously studied by Baker et al. in these groups possessing a ZCWPW1 orthologue. For the latter two, we did find potential orthologues (e.g. in canids), but with

mutations at positions in ZCWPW1 that are conserved among all PRDM9-SET-possessing species (methods section 3.2). Thus, there has been similar (co)evolution of presence/absence of ZCWPW1 and PRDM9.

We observe a greater number of species - 28 - possessing PRDM9 orthologues but having no relative closer than their class possessing an identified ZCWPW1 orthologue. All of these are bony fish. Strikingly, 24 of these 28 orthologues possess one or more mutations in their SET domain, predicted to disrupt H3K4me3 and/or H3K36me3 deposition. In strong contrast, 94% of PRDM9-containing species with non-mutant SET domains have a closer relative possessing ZCWPW1 (odds ratio = 90; indicative $p < 10^{-15}$ by FET, though we note these observations are not all independent). This implies that ZCWPW1 is most often lost, whenever mutations occur in the SET domain of PRDM9. Interestingly, PRDM9-containing species with non-mutant SET domains are near-identical to those also possessing SSXRD domains, and SSXRD has been reported as essential for PRDM9's H3K4me3 methyltransferase activity at hotspots in vitro (Thibault-Sennett et al., 2018). However, other domains of PRDM9 (KRAB, and the zinc finger array) have been lost across multiple species, and it is hypothesized PRDM9 does not position recombination in these species (Z. Baker et al., 2017). Nonetheless, they mainly retain ZCWPW1 . In conclusion, among vertebrates at least, it appears the set of species with ZCWPW1 is extremely similar to those species possessing PRDM9 with both an intact SET domain, and an SSXRD domain, with considerable evidence of co-evolution of gain/loss events for each protein. We note that this pattern is precisely what would be expected a priori, if ZCWPW1's main function involves recognition of the histone modifications catalysed by PRDM9-SET during early meiosis, as predicted from functional considerations.

3.3 ZCWPW1 localisation and KO phenotypes

To investigate the role of ZCWPW1 during meiosis in vivo, we produced an antibody against the full-length recombinant mouse protein and studied the phenotype of a

newly generated knockout (KO) mouse line for Zcwpw1, with a particular focus on fertility and meiotic recombination.

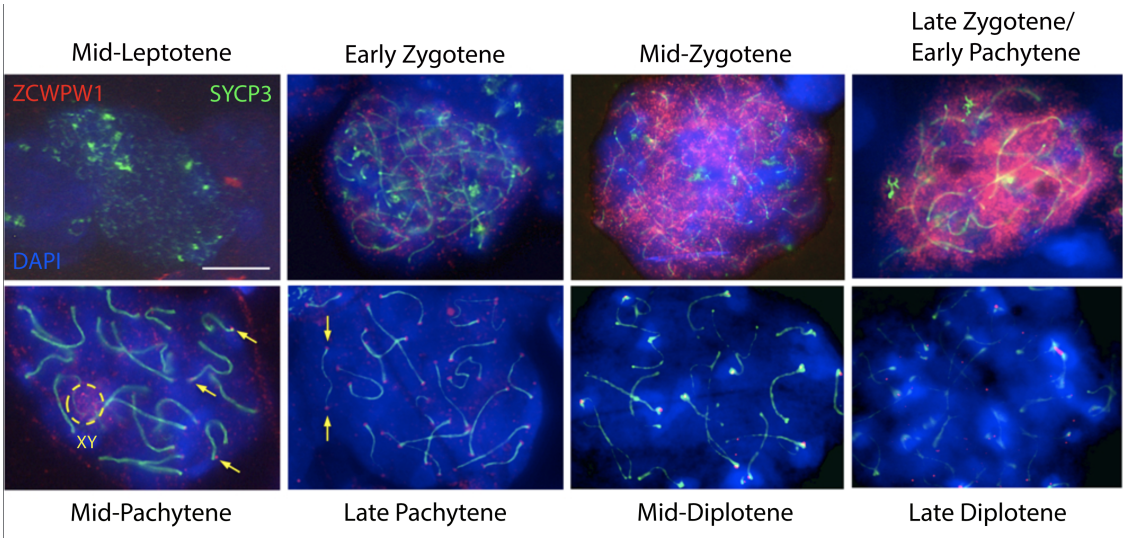


Figure 3.2: Expression of ZCWPW1 across meiosis prophase I in mouse testis. Nuclear spreads were immunostained with antibodies against ZCWPW1 (red) and the synaptonemal complex protein SYCP3 (green) which labels the chromosome axis, and counterstained with DAPI (blue) to visualise nuclei. Developmental stages indicated above and below. Yellow arrows show apparent ZCWPW1 foci at both ends of the synaptonemal complex. The dashed circle shows staining in the XY body. Scale bar: 10 µm Imaging by Daniela Moralli.

In testes from wild-type (WT) mice, we observe a dynamic localisation of ZCWPW1 protein (Figure 3.2), similar but non-identical to that reported in a recent study (M. Li et al., 2019). ZCWPW1 shows a strong, punctate nuclear staining excluding the pericentromeric regions (clustered into chromocenters brightly stained with DAPI) in zygotene and early pachytene cells; we detected transcript expression in earlier pre-leptotene to leptotene cells (Jung et al., 2019). In pachytene cells, ZCWPW1 expression drops, with the protein now mainly localised in the XY body and as bright foci at the ends of the synaptonemal complex labelled by SYCP3, not previously observed using an antibody raised against a 174 base-pair (bp) C-terminal region of the protein (M. Li et al., 2019). By diplotene, little expression is visible. Using FISH to label telomeric and centromeric ends of chromosomes, we established that these discrete foci of ZCWPW1 are located at the ends of the synaptonemal complex (figure A.2).

We next studied mice from a constitutive *Zcwpw1* KO line (methods section 3.8.3), carrying a ~1.5kb frameshift deletion encompassing exons 5 to 7 upstream of the CW domain, creating a premature stop codon resulting in the production and predicted degradation (by RNA-mediated decay) of a short 492bp (vs 1893bp for WT) transcript (figure 3.3A). Confirming this, ZCWPW1 expression was completely absent in testis protein extracts and chromosome spreads from *Zcwpw1*$^{-/-}$ mice, in zygotene cells (Figure 3.3B) and all other meiotic stages of prophase I where expression is detected in WT mice . Confirming recent findings in a different *Zcwpw1*$^{-/-}$ mouse (M. Li et al., 2019), we observed no overt fertility phenotype in either sex in the heterozygous *Zcwpw1*$^{+/-}$ mice (data not shown). However *Zcwpw1*$^{-/-}$ male mice were sterile with complete azoospermia and reduced testis size (Figure 3.3B,D), while female mice retained fertility until around 7-8 months of age , and otherwise both sexes develop normally. As in M. Li et al., 2019, we observe widespread asynapsed chromosomes in male mice, marked by γH2AX and HORMAD2, persistent DMC1 foci marking unrepaired DSBs (Figure 3.4), no meiotic progression beyond (pseudo)pachytene, failure to form the sex body, and a complete absence of MLH1 foci marking recombination crossover sites (figure A.3).

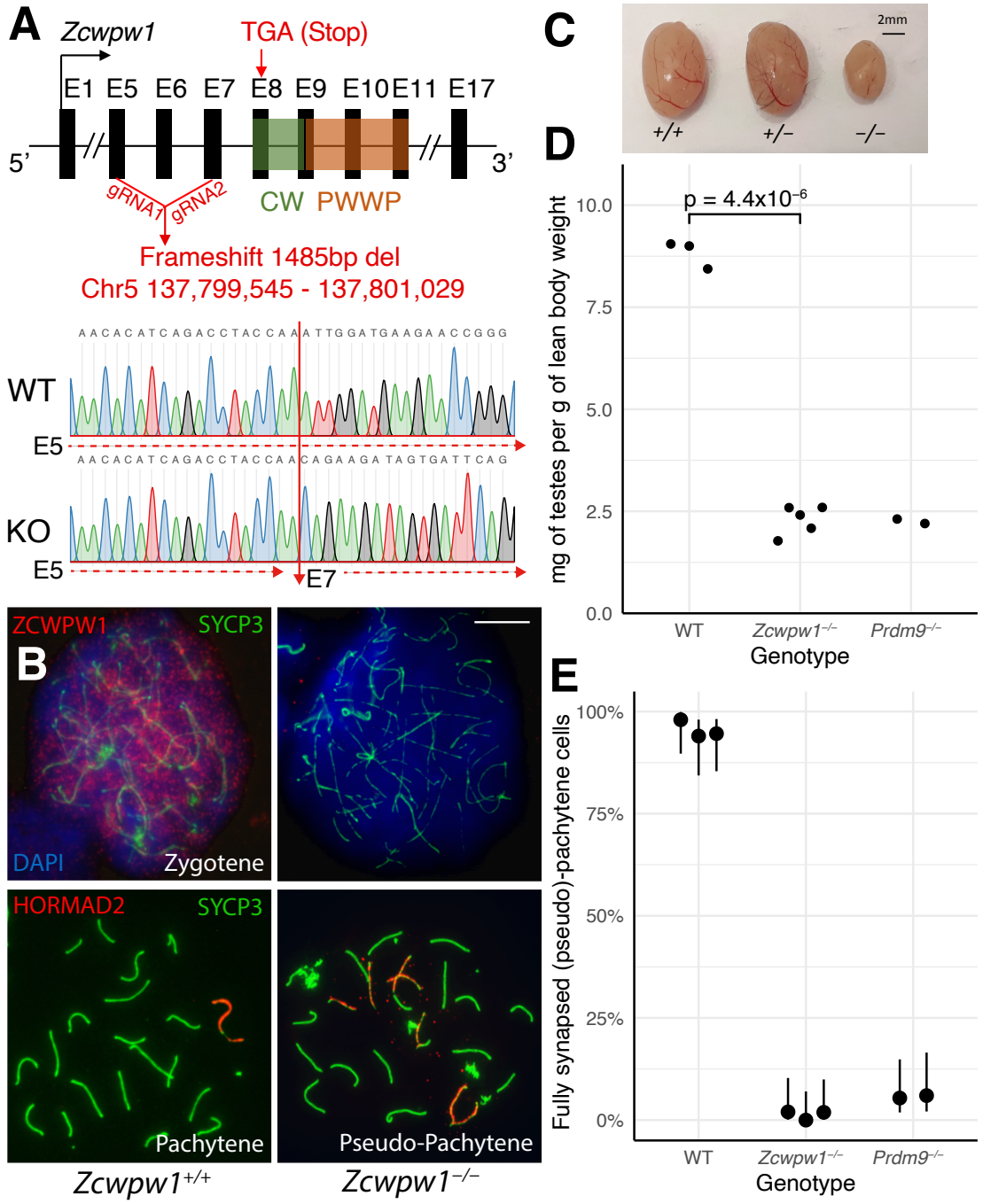


Figure 3.3: Male *Zcwpw1*$^{-/-}$ mice show reduced testis size and asynapsis, similar to the *Prdm9*$^{-/-}$ mutant. **(A)** Schematic of the Zcwpw1 knockout mouse line. E: Exon. gRNA guideRNA. Sanger sequencing DNA chromatograms of wild-type (WT) and knockout (KO) mice encompassing the deletion are shown. The intron-exon organization is not to scale. **(B)** Immunofluorescence staining of testis nuclear spreads from *Zcwpw1*$^{+/+}$ and *Zcwpw1*$^{-/-}$ mice for ZCWPW1, the synaptonemal complex protein SYCP3 which labels the chromosome axis, or HORMAD2 which marks unsynapsed chromosomes. **(C)** Representative testes from 9-10 weeks old WT (+/+), Het (+/-) and Hom (-/-) *Zcwpw1* KO mice are shown. **(D)** Paired testes weight was normalized to lean body weight. The p-value is from Welch's two sided, two sample t-test. **(E)** Synapsis quantification in testis chromosome spreads immunostained with HORMAD2, as in (B). The percentage of pachytene cells with all autosomes fully synapsed is plotted by genotype; $n \geq 50$ cells. Vertical lines are 95% Wilson binomial confidence intervals.

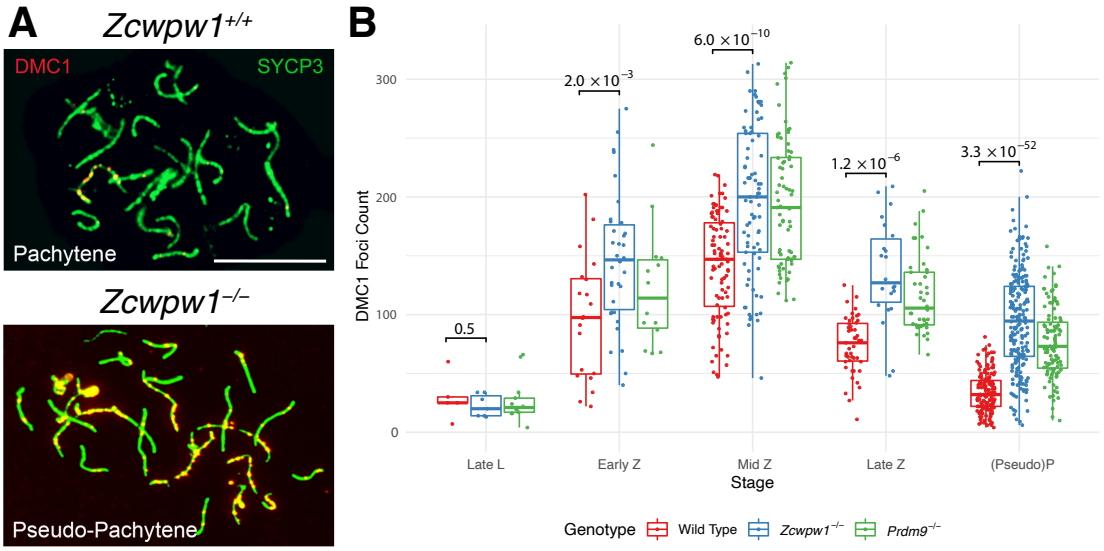


Figure 3.4: Similar DMC1 count elevation in *Zcwpw1*$^{-/-}$ and *Prdm9*$^{-/-}$ mice, compared to wild-type. **(A)** Testis chromosome spreads from wild-type and *Zcwpw1*$^{-/-}$ mice were immunostained for DMC1 and SYCP3. Representative late pachytene cells are shown. **(B)** The number of DMC1 foci in cells from the various stages of prophase I were counted. p-values are from Welch's two sided, two sample t-test. L: Leptotene, Z: Zygotene, P: Pachytene. n=3 mice per genotype (*Zcwpw1*$^{-/-}$ and WT), n=2 for *Prdm9*$^{-/-}$.

Each of these properties resembles observations in the *Prdm9*$^{-/-}$ mutant, and so we compared to this mutant. In our *Zcwpw1*$^{-/-}$ male mice, >98% of pachytene cells failed to properly synapse at least one pair of chromosomes (Figure 3.3B, E), similar to *Prdm9*$^{-/-}$ males. However, the nature of the synaptic defects observed differed. In the *Prdm9*$^{-/-}$ mutant, 63.2% of the pseudopachytene cells showed mispairing of non-homologous chromosomes in a typical branched structure. In contrast, we only observed 24.8% of *Zcwpw1*$^{-/-}$ pseudopachytene cells with this type of error, while the majority (74.5%) of cells contained multiple bundles of HORMAD2-positive unsynapsed chromosomes, which resemble the XY body and may merge with the sex chromosomes. These results imply that the *Prdm9*$^{-/-}$ mutant often mispairs chromosomes, while *Zcwpw1*$^{-/-}$ spermatocytes mainly fail to pair a subset of chromosomes at all. The expression levels and staining pattern of ZCWPW1 were not visibly altered in testis nuclear spreads (data not shown) and protein extracts from *Prdm9*$^{-/-}$ mice.

Comparing levels of DMC1 foci as a proxy for DSB repair, or a potential cause for the asynapsis in *Zcwpw1*$^{-/-}$ males, in both *Zcwpw1*$^{-/-}$ and *Prdm9*$^{-/-}$ mice foci count was significantly elevated from early zygotene onwards, indicating delayed repair of DSBs (Figure 3.4). However, we observed a wider spread for *Zcwpw1*$^{-/-}$ males. A similar increase was observed at pachytene stage in the levels of RAD51 (figure A.4), another ssDNA-binding protein which functions in concert with DMC1, with the large majority of RAD51 and DMC1 foci co-localizing (Brown, Grubb, et al., 2015; Tarsounas et al., 1999). In *Zcwpw1*$^{-/-}$ males, like in the *Prdm9*$^{-/-}$ mutant, DSBs form and recruit RPA2, RAD51 and DMC1 in similar numbers (M. Li et al., 2019), but fail to repair efficiently, accompanied by asynapsis and meiotic arrest at pseudo-pachytene. Indeed, we observe late unrepaired DMC1 foci mainly on asynapsed chromosomes.

3.4 ZCWPW1 recruitment by PRDM9

We previously studied the binding properties of human PRDM9 and established a genome-wide map in transfected human mitotic (HEK293T) cells by ChIP-seq (Altemose et al., 2017), observing binding to the majority of human meiotic recombination hotspots. Based on the presence of H3K4me3 and H3K36me3 recognition domains in ZCWPW1, we hypothesized that it would be recruited to PRDM9-bound genomic sites, where these marks are deposited upon binding in HEK293T cells (Altemose et al., 2017).

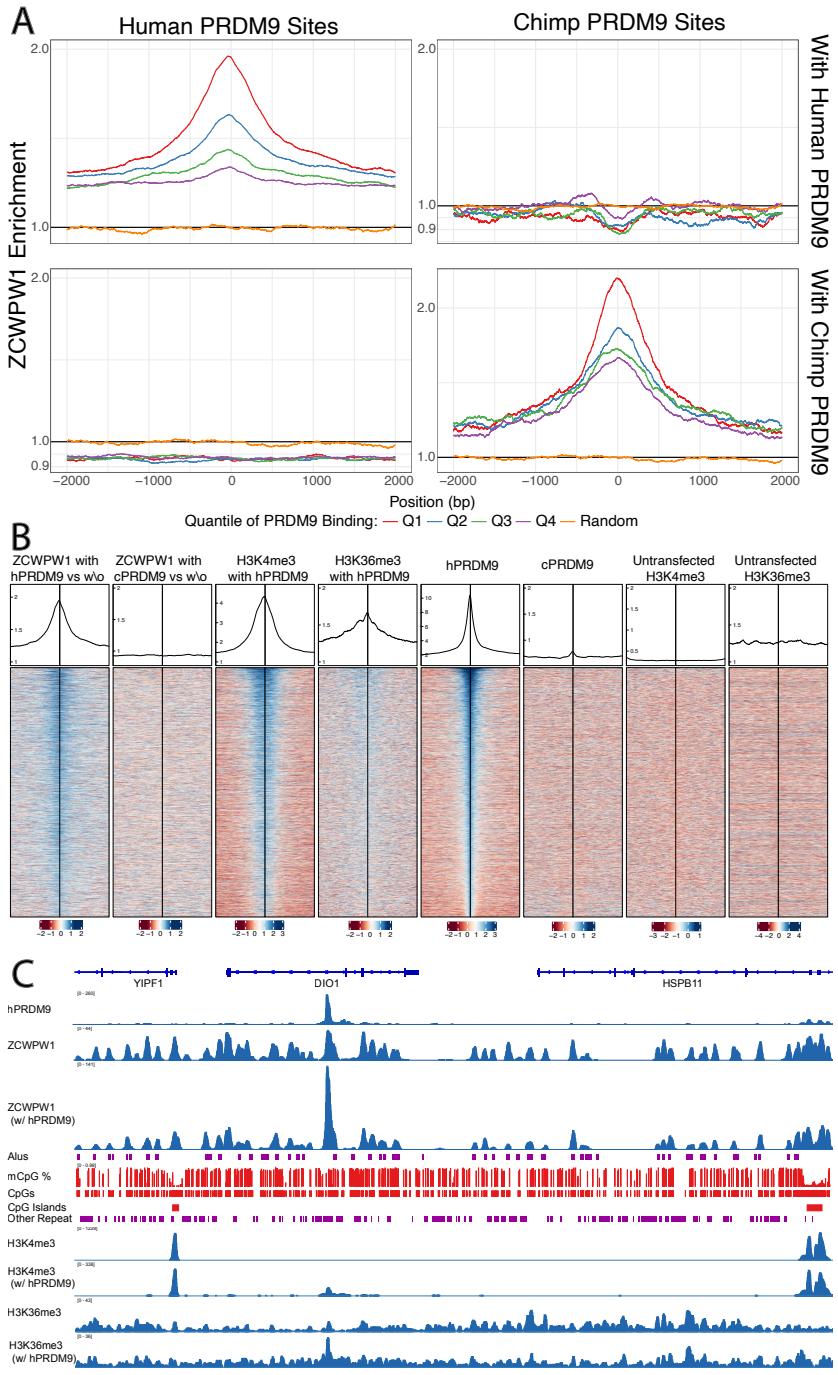


Figure 3.5: **(A)** Enrichment of ZCWPW1 (with vs without PRDM9) at PRDM9 binding sites when co-transfected with PRDM9 with either Human or Chimp Zinc Finger. Q = quartile. Human PRDM9 sites are centered and stranded by the motif. **(B)** Profiles and heatmaps of reads around top 25% of individual human (h) PRDM9 binding sites (rows). Heatmaps: log-fold change of target (indicated in column titles, methods section 3.8.10) vs input, for various labeled target proteins, ordered by human PRDM9. ZCWPW1, H3K4me3 and H3K36me3 each become enriched at human PRDM9 sites, following (co-)transfection with human PRDM9. **(C)** ChIP-seq data and annotation in a genome plot illustrate the behaviour of ZCWPW1 and other factors. ChIP-seq tracks show fragment coverage. Tracks where PRDM9 is present are labelled "w/ PRDM9", and below corresponding tracks with out PRDM9. ZCWPW1 binds to Alus, CpG islands and other CpG-rich sequences even in the absence of PRDM9. On addition of PRDM9, ZCWPW1 becomes strongly enriched at PRDM9 binding locations (center left peak within DIO1).

To test this, we co-transfected HEK293T cells with full-length human HA-tagged ZCWPW1 and either no other protein, or full-length PRDM9 alleles carrying the human or chimpanzee ZF-array, as studied previously (Altemose et al., 2017), and then performed ChIP-seq against the ZCWPW1 HA tag. Confirming the recruitment hypothesis, in the presence of human PRDM9 and compared to cells lacking human PRDM9, ZCWPW1 shows a strong enrichment at human PRDM9 binding sites with higher enrichment at sites with higher PRDM9 enrichment (figure 3.5A, B). Notably, even without PRDM9 we observed sequence-specific binding of ZCWPW1 to many sites in the genome, some coinciding with PRDM9 binding sites (Figure 3.5C and A.5). However, upon transfection, 92% (98%) of the strongest 10,000 (3,000) peaks are PRDM9 binding sites (cf. 7% expected overlap with randomised peaks), so PRDM9 is able to strongly reprogram ZCWPW1 binding, suggesting that non-PRDM9 peaks are bound more weakly. Because our transfection has <100% efficiency (figure A.6), some cells containing ZCWPW1 will not possess PRDM9, and so it is possible that an even greater fraction of ZCWPW1 is redirected to PRDM9 binding sites in cells where both proteins are present.

Importantly, when co-transfecting with a modified version of PRDM9 in which the zinc finger array is replaced with that from chimp (which binds different locations in the genome (Altemose et al., 2017)), we find that the enrichment at human binding sites disappears, but instead ZCWPW1 is enriched at chimp PRDM9 binding sites (Figure 3.5A and A.7). This perturbation experiment provides strong evidence that PRDM9 causes recruitment of ZCWPW1 (as opposed to for example independent recruitment of both proteins). Notably, the strength of binding of ZCWPW1 at these sites provides a better predictor of DSB formation (DMC1 sites) than does PRDM9 binding strength itself (figure A.8): although we show (see below) that ZCWPW1 is not directly involved in DSB positioning, this might suggest involvement of similar features to those it recognizes, in recruiting DSBs formation machinery.

We tested whether this recruitment might be mediated by the dual histone modifications H3K4me3 and H3K36me3. Consistent with this idea, ZCWPW1 binding is positively associated with levels of both H3K4me3 and H3K36me3 marks

(figures A.9 and A.10). We examined transcription start sites, which possess H3K4me3 at high levels, but lack H3K36me3, observing some ZCWPW1 signal at these sites, but with a uniformly lower mean ZCWPW1 enrichment compared to those with evidence of PRDM9 binding (and hence both histone marks) (Figure 3.6 and A.11). Although it is not possible to measure H3K4me3 presence/absence in individual cells in our system, even the most weakly bound PRDM9 sites – which must therefore possess H3K4me3 in only a fraction of cells – show stronger ZCWPW1 enrichment than the strongest promoters, which are likely to have near 100% H3K4me3 marking, and possess $>$ 2-fold more H3K4me3 than even the strongest PRDM9 binding sites. We conclude that H3K4me3 alone endows only relatively weak binding, while PRDM9 is therefore able to recruit ZCWPW1 with a much greater efficiency than sites marked by H3K4me3 alone, suggesting that both histone modifications might aid efficient binding.

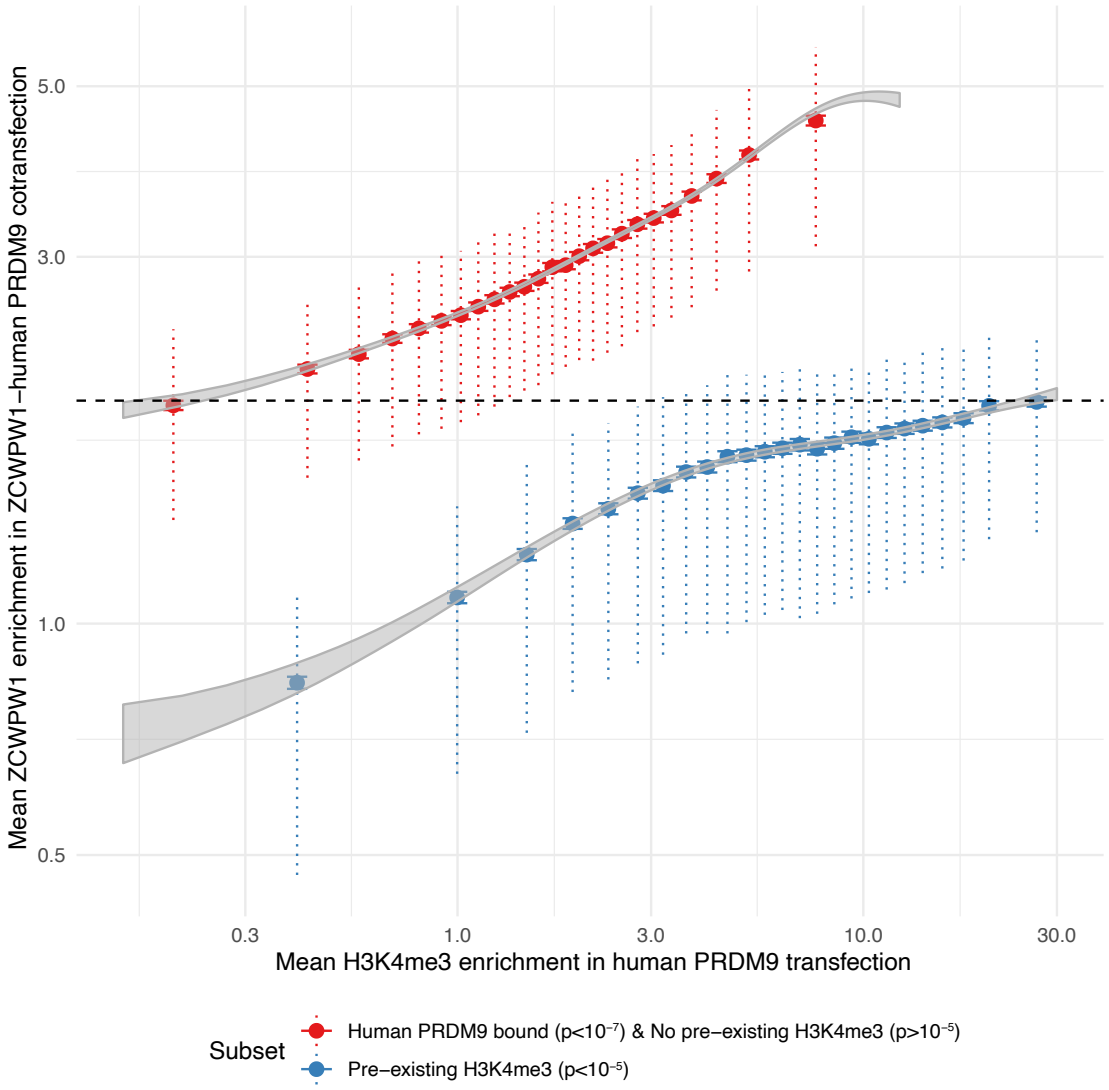


Figure 3.6: PRDM9-bound regions (H3K4me3 and H3K36me3) are a stronger recruiter of ZCWPW1 than promoters (H3K4me3 only). For any given level of H3K4me3 (x-axis), ZCWPW1 enrichment (y-axis) is higher at PRDM9-bound regions (red) than regions with pre-existing H3K4me3 (promoters, blue). H3K4me3 and ZCWPW1 were force called in 100bp windows. Each subset is defined as indicated in the legend with the additional constraint of requiring input fragment coverage >5 for ZCWPW1 and >15 for H3K4me3. For each subset H3K4me3 was split into 25 bins with equal number of data points. Horizontal bars: 2 standard errors of the mean. Vertical dotted bars: upper and lower quartiles. Grey ribbons show 2 standard errors for a Generalized Additive Model on log(mean H3K4me3 enrichment + 0.1). Dashed black horizontal line highlights that the mean enrichment of the highest bin for promoters is similar to that of the lowest bin for PRDM9 bound sites.

3.5 DSBs positioning and repair

Previous work has shown that *Prdm9*$^{-/-}$ mice use a new set of DSB hotspots, localising at CpG islands and/or promoter regions (Brick et al., 2012). Given that PRDM9 recruits ZCWPW1, one possible function of ZCWPW1 may be that, in turn, it recruits the DSB machinery and hence forms part of the causal chain in normal positioning of DSBs at PRDM9-specified hotspots. Alternatively, the *Zcwpw1*$^{-/-}$ mutant phenotypes we observe might reflect a more downstream role. To distinguish these hypotheses, we gathered data on DSB positioning and repair dynamics genome-wide, by carrying out single stranded DNA (ssDNA) sequencing (SSDS) by ChIP-seq against DMC1 (Khil et al., 2012).

In the *Zcwpw1*$^{-/-}$ mutant, we observed normal localisation of DMC1 at the expected B6 hotspots for this background, with no activity at *Prdm9*$^{-/-}$ hotspots (Figure 3.7A). Moreover, individual wild-type hotspots appear always to be active in the *Zcwpw1*$^{-/-}$ mutant (figure 3.7C): we only fail to see evidence of DMC1 signal in a subset of the weakest hotspots, where we are likely to lack statistical power (figure A.12). Thus DSBs occur in unchanged hotspot regions in *Zcwpw1*$^{-/-}$ males. To check if they occur at the same locations within hotspot regions, we leveraged data for SPO11-mapped DSB sites (Lange et al., 2016). Specifically, we identified three sets of hotspots whose respective mapped breaks occur mainly upstream, central, or downstream of the PRDM9 binding site (Methods), and compared their DMC1 ChIP-Seq signal profiles, as well as their SPO11 signals (figure A.13). This revealed that in the *Zcwpw1*$^{-/-}$ males, as in the wild-type, DMC1 signals, mirror this break positioning. Therefore, ZCWPW1 is not required to specify hotspot locations, and neither does it strongly influence where DSBs occur within hotspots.

One perhaps important difference between the *Zcwpw1*$^{-/-}$ and wild-type DMC1 profiles is that the latter signal is slightly wider (up to ~200bp) in hotspots with increasing PRDM9-induced H3K4me3 (Figure 3.7B), while this effect is absent in the *Zcwpw1*$^{-/-}$ case. This might be explained by small differences in chromatin accessibility subtly impacting SPO11 locations, DNA end resection which generates

the 3' ssDNA tail to which DMC1 binds, or downstream processing differences altering DMC1 span in the mutant mice.

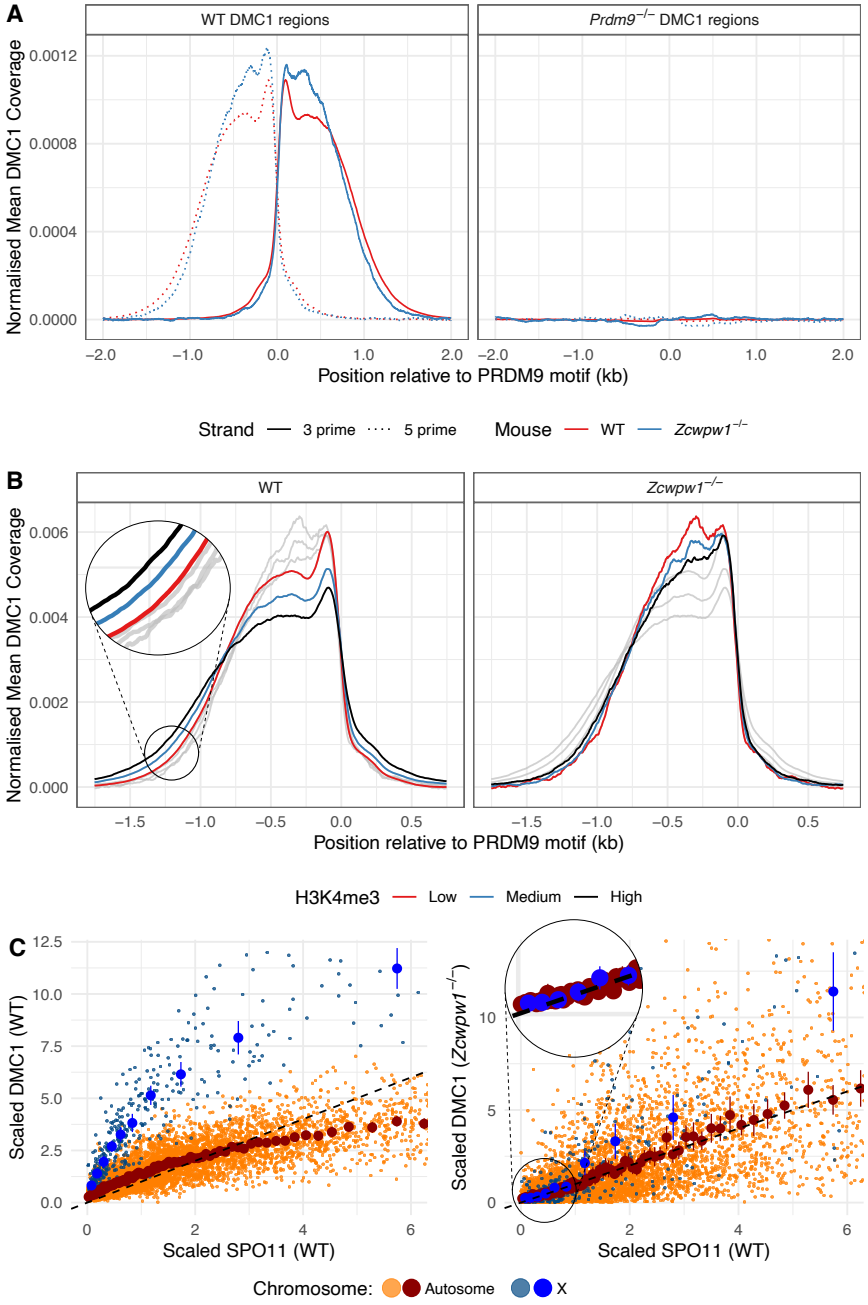


Figure 3.7: **(A)** DSBs occur at normal hotspot locations in the *Zcwpw1*$^{-/-}$ mouse. Average coverage of reads from DMC1 SSDS ChIP-seq at previously mapped regions (methods section 3.8.10) in B6 WT (left) and *Prdm9*$^{-/-}$ (right) mice is shown, centered at the PRDM9 motif (left). DMC1 profiles from a WT mouse are shown in red, data from (Brick et al., 2012). **(B)** Normalised DMC1 profile (both strands combined) is plotted for WT and *Zcwpw1*$^{-/-}$, stratified by H3K4me3 (a proxy for PRDM9 binding). Low: <50th percentile cumulative enrichment, High: >75th percentile cumulative enrichment, with Medium being the remaining data. Dashed lines show the alternative genotype for comparison. **(C)** Altered DMC1 signals in *Zcwpw1*$^{-/-}$ mice. Relationship between WT SPO11-oligos (measuring the number of DSBs) vs DMC1 (a measure of the number and persistence of DSBs) at each B6 hotspot for WT and *Zcwpw1*$^{-/-}$. Unlike WT mice, DMC1 signals in *Zcwpw1*$^{-/-}$ mice are approximately linearly associated with WT SPO11. Black dashed line is y = x for reference. SPO11 and DMC1 enrichment have been scaled by dividing by the mean autosomal enrichment. Large dark blue and dark red points show mean DMC1 signal, binned into groups containing equal numbers of hotspots by WT SPO11 signal, for X (10 bins) and autosomal data (100 bins) respectively (smaller lighter dots represent individual hotspots).

Despite very similar DSB locations, we saw much greater, and systematic, differences in the strength of the DMC1 signal at individual hotspots. As previously (B. Davies et al., 2016; Khil et al., 2012), we note that observed average DMC1 signal strength at a hotspot reflects the product of the frequency at which DSBs occur there, and the average length of time DMC1 remains bound to the ssDNA repair intermediate, with the latter reflecting DSB processing/repair time. In contrast, available SPO11 ChIP-Seq data (Lange et al., 2016) reflect mainly the former, the frequency of DSBs. We therefore compared DMC1 and SPO11 signal strength at each autosomal and X-chromosome hotspot, in WT and *Zcwpw1*$^{-/-}$ mice. In WT mice (as seen previously in other mice (B. Davies et al., 2016) including, albeit somewhat more weakly, even sterile hybrids) the non-PAR X-chromosome shows a very strong elevation of DMC1 signal strength, reflecting the persistent DMC1 foci on this chromosome also visible using microscopy, at DSB sites that eventually repair using the sister chromatid. Moreover, in WT mice hotspot heat (with hotter hotspots having stronger H3K4me3 signal also) shows a sub-linear relationship between SPO11 and DMC1. This is thought to reflect a wider phenomenon of faster DSB repair occurring within those hotspots whose homologue is more strongly bound by PRDM9, i.e. those hotspots with a stronger H3K4me3 signal in the WT mouse (B. Davies et al., 2016; Hinch et al., 2019; R. Li et al., 2019). The X-chromosome DMC1 elevation is in a sense an extreme case of slower repair of DSBs whose homologue is not PRDM9-bound, because no homologue exists in this case.

However, we see a striking departure in the *Zcwpw1*$^{-/-}$ mouse, where the X-chromosome behaves similarly to the autosomes in terms of DMC1 vs. SPO11 signal strength (figure 3.7C). Moreover, we also see a simple linear relationship between DMC1 and SPO11 binding in this mouse. These data imply the relationship between DMC1 removal and PRDM9 binding is eliminated in this mouse, so PRDM9 (whether because ZCWPW1 directly aids the process, or indirectly) appears unable to "assist" in homolog pairing and synapsis in this mouse. Indeed, the data imply a widespread perturbation of DSB repair in this mouse, with autosomal DMC1 foci persisting as long as those on the X-chromosome. In the previously studied

Hop2$^{-/-}$ mouse (Khil et al., 2012; Petukhova et al., 2003; Smagulova et al., 2011), we identified a very similar pattern (figure A.14): in this mouse, DMC1 is loaded onto ssDNA, but DSBs are not repaired at all, indicating that DSB repair pathways in *Zcwpw1*$^{-/-}$ mice are profoundly altered.

These results are consistent with, and extend, our microscopy observations (figure 3.4) that many DSBs persist in the *Zcwpw1*$^{-/-}$ mouse, and some may never repair. We are unable to say whether DSB repair involves the homologue or not in this mouse, but the partial synapsis we observe suggests some repair does likely occur.

To further understand DMC1 persistence at individual hotspots, we estimated the relative DMC1 heat of autosomal hotspots in the *Zcwpw1*$^{-/-}$ mouse, compared to the WT mouse. The best individual predictor of this ratio among H3K4me3, WT DMC1, and SPO11 was the level of H3K4me3 ($r = 0.60$, Generalised Additive Model, figure A.15, while $r = 0.63$ when using all 3 predictors together). The average DMC1 ratio changes around 9-fold from the most weakly to the most strongly H3K4me3-marked hotspots. Therefore, if DMC1 signal changes are indeed explained by slower DSB processing in the *Zcwpw1*$^{-/-}$ mouse at hotspots bound strongly by PRDM9, then this implies a very strong effect of PRDM9's "assistance" in aiding repair. Whatever the cause, the overall strong correlation implies perturbation of DMC1 behaviour is widespread, rather than impacting any small subset of hotspots. Moreover, it appears to be mainly controlled by local levels of PRDM9 histone modification, in keeping with the evidence that ZCWPW1 recognizes this mark at PRDM9-bound sites, on either the broken or (identical) homologous chromosome.

3.6 ZCWPW1 binds CpG dinucleotides

In addition to the strong PRDM9-dependent ZCWPW1 peaks described earlier, there are many locations in HEK293T cells at which ZCWPW1 binds, typically more weakly, and independently of PRDM9. Indeed we identified over 800,000 ZCWPW1 peaks. Surprisingly, a large proportion of these binding sites overlap Alu repeats (Figure 3.8A and A.16) (of which there are 1.1 million in the human genome (Deininger, 2011)). The weakest ZCWPW1 peaks overlap Alus most

frequently, whilst the strongest peaks are depleted of Alus relative to chance overlap (Figure 3.8A).

Because Alus are rich in CpGs (containing >23% of human CpG dinucleotides, with deamination depleting CpGs in older Alus (Luo et al., 2014)), we tested if binding correlates with CpG presence/absence. Indeed, the binding of ZCWPW1 to Alus depends on the presence of CpGs, which are mainly methylated (Gaysinskaya, B. F. Miller, et al., 2018): Alus with no CpGs were bound at a rate lower than by chance, but almost all Alus containing 10-20 CpGs are bound (Figure 3.8B). This dependence is not Alu-specific: even outside genomic repeats, 300-bp regions with zero CpGs have negligible probability of overlapping a ZCWPW1 binding site, while the overlap probability rises to >50% for regions containing 10 or more CpGs, suggesting CpGs are essential for PRDM9-independent ZCWPW1 binding (figure 3.8C,D and A.17). ZCWPW1 appears to have greater affinity for methylated CpG pairs, but retains some affinity even for non-methylated regions. This binding mode was unexpected a priori, and does not appear to be easily explained by patterns of H3K4me3/H3K36me3 in the genome, which do not concentrate as strongly in these regions.

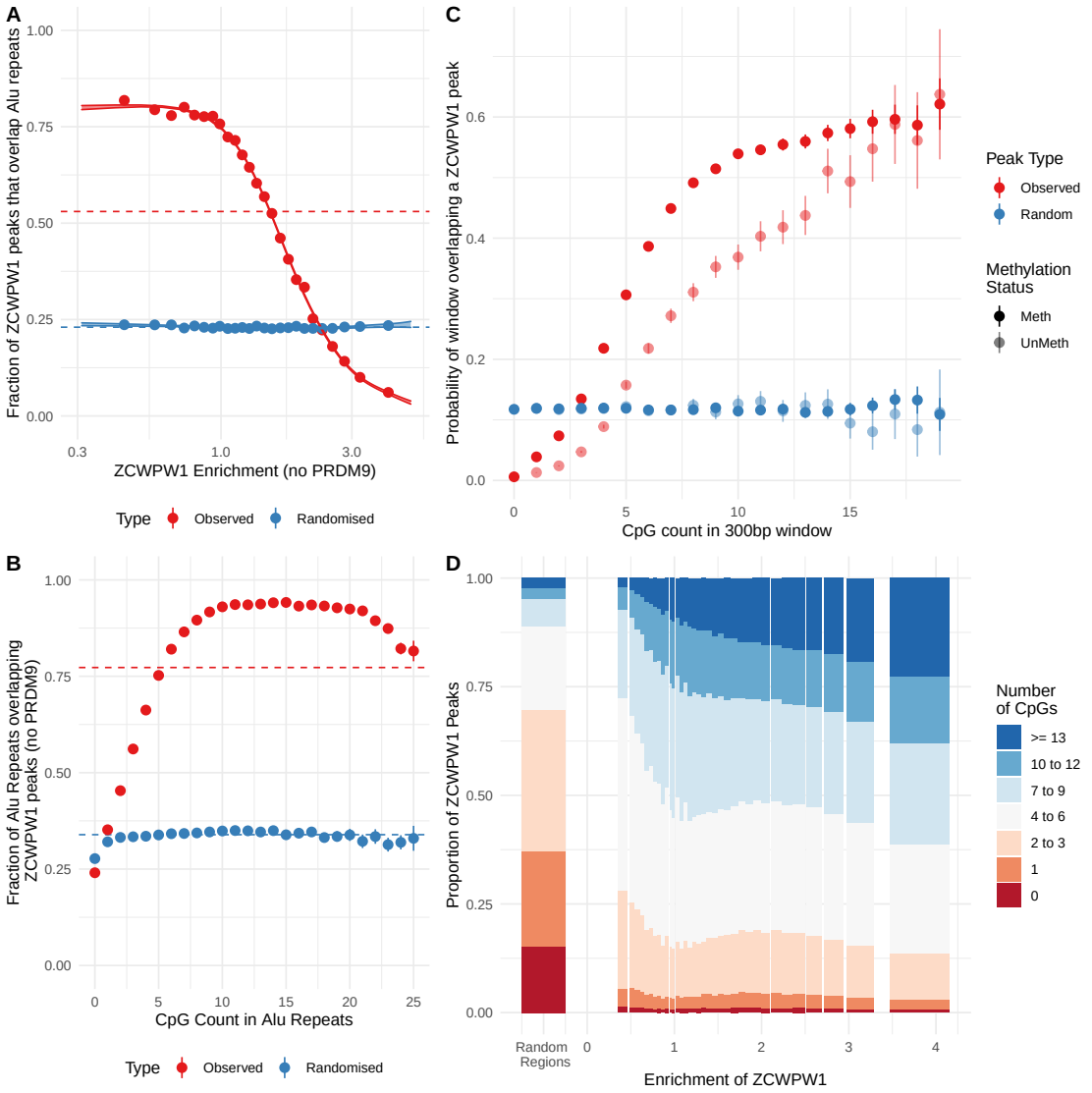


Figure 3.8: ZCWPW1 binds CpG-rich sequences such as Alu repeats. **(A)** Fraction of overlap of ZCWPW1 binding peaks with Alus, ordered by enrichment in ZCWPW1 binding. ZCWPW1 peaks are binned into 25 bins with equal number of data points, and means of both enrichment and overlap are plotted. Solid ribbon shows prediction from GAM logistic model. Dotted line shows overall means. **(B)** Rate of overlap of Alu repeats with ZCWPW1 peaks, for Alus with different numbers of CpG dinucleotides. **(C)** The probability of a 300bp window overlapping a ZCWPW1 peak increases with increasing CpG count in that window. Windows overlapping (by 10bp or more) Alus, other repeats, or CpG islands have been excluded. Methylated regions are those with a methylated to unmethylated reads ratio of >0.75, and unmethylated <0.25 (methods section 3.8.10) **(D)** Relative proportion of peaks with given numbers of CpGs ±150bp from peak center. ZCWPW1 peaks are enriched in CpGs compared to random peak locations. 'Meth', methylated; 'UnMeth', unmethylated.

3.7 Discussion

Here we exploited previous work on single-cell transcript profiling of mouse testis, which has established spatio-temporal (and likely functional) components of co-expressed genes (Jung et al., 2019), to identify co-factors relevant to PRDM9 function in specification of recombination hotspots and DSB repair during meiosis. We show that the histone methylation reader ZCWPW1, which is both highly co-expressed and co-evolving with PRDM9, is recruited by PRDM9 to its binding sites, likely mediated through the recognition of the dual H3K4me3-H3K36me3 histone mark PRDM9 deposits.

While we find co-evolution of ZCWPW1 with PRDM9, consistent with proteins functioning in the same pathway, this appears to depend on particular domains. ZCWPW1 shows weaker conservation among species outside the CW and PWWP domains, suggesting these are of high importance, and co-occurs with PRDM9 in species where the latter possesses a SET domain predicted to retain catalytic activity for H3K4me3/H3K36me3, which the domains of ZCWPW1 are predicted to recognise, strongly suggesting a partnership totally dependent on this recognition. This co-occurrence can persist, even in species where PRDM9 is not believed to position DSBs, suggesting the co-operation between these proteins does not completely depend on this aspect of PRDM9's function. Interestingly, such a SET domain of PRDM9 mainly occurs when PRDM9 possesses a SSXRD domain, which in other proteins (Banito et al., 2018) recruits CXXC2 which binds to CpG islands – whether this connects to our observation of a potential CpG-binding domain in ZCWPW1, and CpG-binding behaviour in a human cell line, remains to be explored.

A recent study reported phenotyping of an independent *Zcwpw1*$^{-/-}$ mouse (M. Li et al., 2019), and we observe similar patterns of infertility in males, and results consistent with gradual fertility decline in females, towards sterility from 8 months onwards. Why females deficient for ZCWPW1 are (at least initially) fertile is unknown, but resembles sexual dimorphism in loss-of-function mutants for other meiotic genes involved in homologous recombination, where females exhibit a milder fertility phenotype (Cahoon et al., 2019; Morelli et al., 2005; J. Zhang et al., 2019).

However, we observe a distinct intracellular localisation of ZCWPW1 from that reported (M. Li et al., 2019): while we also detect ZCWPW1 as a diffuse nuclear signal across the nucleus (though excluding the chromocenters) and localisation to the XY body in pachytene, we see an additional signal in mid-pachytene to diplotene cells at the ends of the synaptonemal complex. This might reflect the specificity of our polyclonal antibody, which is raised against the full-length mouse protein vs a monoclonal antibody raised against a C-terminal region, hence recognising a wider spectrum of ZCWPW1. Given that the *Zcwpw1*$^{-/-}$ mouse arrests in early pachytene, it is not possible to test later stages, so a caveat is that we cannot completely exclude off-target antibody recognition. If accurate, the late-stage localisation of ZCWPW1 at subtelomeric regions resembles other proteins (Speedy A, CDK2, TFR1, SUN1, TERB1/2, MAJIN, and KASH), which form a complex tethering the telomeres to the nuclear envelope to facilitate chromosome movement and pairing - essential for synapsis and fertility (Ashley et al., 2001; Ding et al., 2007; Horn et al., 2013; Shibuya et al., 2014; Tu et al., 2017). However, no evidence of impaired telomeric attachment was reported in the other *Zcwpw1*$^{-/-}$, based on normal localisation of TRF1 (M. Li et al., 2019).

We find, using a model system, that ZCWPW1 is strongly recruited to PRDM9-bound sites, in an allele-specific manner, and dependent on the level of H3K4me3 and H3K36me3 that PRDM9 deposits. Hence recruitment is most likely due to recognition of these marks by the CW and PWWP domains of ZCWPW1. Li and colleagues (M. Li et al., 2019) showed co-immunoprecipitation of H3K4me3 with ZCWPW1, but our results imply that PRDM9 recruits ZCWPW1 much more strongly than H3K4me3 alone, further suggesting that the combination of histone marks is key for ZCWPW1's recognition of PRDM9 sites. Indeed in cells with PRDM9, the strongest ZCWPW1 binding sites are almost all PRDM9-bound sites. To the best of our knowledge, this is the first protein identified that directly recognises PRDM9 binding sites.

Although it has been suggested that ZCWPW1 might recruit the DSB machinery to PRDM9 binding sites (M. Li et al., 2019; Spruce et al., 2019), we find that

DSB formation and positioning are completely unaffected by loss of ZCWPW1. However, the additional role PRDM9 plays besides DSB positioning, in enabling homologous chromosome pairing (B. Davies et al., 2016), malfunctions in *Zcwpw1*$^{-/-}$ mice. Unlike in wild-type animals, DSBs at sites bound more strongly by PRDM9 do not repair more quickly. Instead, in *Zcwpw1*$^{-/-}$ males, all DSBs behave similarly to those on the X-chromosome - which repairs DSBs very slowly, even in wild-type animals, and does not synapse. This implies an unexpected role for ZCWPW1: in aiding PRDM9 in its second function, synapsis. ZCWPW1 therefore offers a (thus far) unique protein, linking PRDM9 binding with its downstream functions (B. Davies et al., 2016; Hinch et al., 2019; R. Li et al., 2019) in aiding rapid DSB processing/repair at hotspots whose homologue is bound by PRDM9.

PRDM9 is thought to aid homology search/chromosome pairing by such binding to the homologous chromosome, and hybrids where hotspots are highly "asymmetric" (i.e. the homologues are not bound at the sites where DSBs occur, due to evolutionary hotspot erosion) show multiple features identical to *Zcwpw1*$^{-/-}$ mice. These include asynapsis (particularly of shorter chromosomes with fewer DSBs), persistent DMC1 foci (at the asymmetric hotspots), and complete sterility in males only. In *Zcwpw1*$^{-/-}$ mice, we infer that the level of PRDM9 binding to the homologue now makes no difference to DSB persistence, so these mice behave, in many ways, as if all hotspots are completely asymmetric. Thus, a parsimonious explanation of our findings is that ZCWPW1 functions by attaching to PRDM9-bound sites on the matching homologous chromosome, thus marking possible repair partner locations for DSB sites, to help pair homologues. This might e.g. lead to recruitment to the chromosomal axis (where DMC1 foci are found), for potential use as the homologous repair template in DSB repair, perhaps mediated by the SCP-1-like domain identified in some orthologous copies. Given we observe ZCWPW1 throughout the nucleus in early meiosis, such recruitment would likely be transient in nature. Alternatively, ZCWPW1's part in mediating PRDM9's downstream roles might also depend (or even, depend only) on it binding the chromosome on which DSBs occur.

Given that we still observe some synapsis, similar to *Prdm9*$^{-/-}$ mice, we suggest that a "back-up" mechanism, which is independent of PRDM9 binding level, is acting to attempt homology search/chromosome pairing in *Zcwpw1*$^{-/-}$ mice. In fact, some such mechanism is likely necessary, to repair DSBs happening within a subset of hotspots that are asymmetric due to naturally occurring variation (as seen in e.g. humans and heterozygous mice).

The simplest explanation of our observation that ZCWPW1 binds in a CpG dependent manner in vitro is that these sites might be directly bound by a methyl-CpG binding domain (MBD) putatively observed in the human, mouse, and coelacanth *Zcwpw1*, among others. Alternatively, recruitment might be indirect, with ZCWPW1 forming a complex with an MBD containing protein, e.g. SETDB1 which is expressed in HEK293 cells, co-expressed with ZCWPW1 and PRDM9 (Jung et al., 2019; D. C. Schultz et al., 2002), and interacts with several other identified PRDM9 partner proteins (Mulligan et al., 2008; Parvanov et al., 2017). Given the precisely regulated developmental changes in methylation status, ZCWPW1/PRDM9 protein abundance, and chromatin during meiosis (Gaysinskaya, B. F. Miller, et al., 2018; Seisenberger Stefanie et al., 2013) it is unclear whether ZCWPW1 will also weakly bind Alus in vivo. However, if CpG dinucleotides do retain weak in vivo ZCWPW1 binding affinity, and given that CpG methylation is associated with transposons, it is interesting to speculate whether ZCWPW1 might play a role in germline transposon recognition, silencing, or facilitate repair of DSBs occurring within transposable elements.

Zcwpw1 possesses a paralogue, *Zcwpw2*, which is also co-expressed in testis with *Prdm9* (although at a lower level), possesses both CW and PWWP domains, and has been shown to be able to recognise H3K4me3 at least (Y. Liu, Tempel, et al., 2016). Further studies are required to investigate what function if any *Zcwpw2* has in meiosis. One interesting possibility is that *Zcwpw2*, rather than *Zcwpw1*, might act to help position DSBs – if so, its low abundance relative to *Zcwpw1* would be analogous to the low number of DSBs per meiosis (~300) relative to the number of PRDM9 binding sites.

In this study we have shown that *Zcwpw1* co-evolves with *Prdm9*, binds to sites marked by PRDM9, and is required for the proper processing, removal of DMC1, and repair of DSBs. How exactly it mediates repair, if for example by recruitment of other proteins, remains to be determined. This study also further demonstrates that genes co-expressed with PRDM9 represent a rich source of undiscovered meiotic genes, with important functional implications for recombination, meiosis progression and ultimately fertility. ZCWPW1 impacts synapsis, downstream of DSB formation. Therefore the protein(s) responsible for recognising PRDM9-bound sites for DSB formation, and for aiding synapsis, are at least partly different, and we anticipate that future studies will likely uncover additional PRDM9-recruited proteins.

3.8 Methods

3.8.1 Orthologue Alignment

We identified ZCWPW1 orthologues across species using four data sources: first, we used BlastP (`blast.ncbi.nlm.nih.gov`), against the full-length human reference ZCWPW1 sequence (identifier NP_060454.3, against nr_v5 database) storing the top 1000 hits using the default parameters (set 1). Secondly, we downloaded the two sets of Ensembl identified ZCWPW1 orthologues (against ENSMUSG00000037108.13; ensembl.org, 105 orthologues), and identified NCBI ZCWPW1 protein orthologues (146 species, www.ncbi.nlm.nih.gov). Initial examination of identified orthologues revealed conservation among orthologues mainly of the ZW and PWWP domains; therefore, to find additional orthologues sequences we performed tBLASTn (against the nr/nt nucleotide collection 17th July 2019; top 1000 hits) to identify orthologues in the NCBI nucleotide database, to the partial sequence (amino acids 256-339 of the reference sequence NP_060454.3), corresponding to the ZW and PWWP domains of ZCWPW1. Protein sequences were then aligned against full-length human ZCWPW1 (NP_060454.3) using BLASTP2.9 (Altschul et al., 1997; Altschul et al., 2005). We obtained taxonomy information from the folder `https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/`, for comparisons among species and between ZCWPW1 and PRDM9.

Previous work has identified highly conserved "KWR" and "PWWP" patterns among CW and PWWP domains, respectively (F. He, Umehara, et al., 2010; Qin et al., 2014). We thus identified an initial set of "clear" ZCWPW1 orthologues, and then used these to identify further less precise matches. "Clear" orthologues are defined as proteins within set 1 containing perfect matches to both these sequences, and such that in the NCBI alignment, at least 39% of each sequence aligns to the human protein, and conversely. This second step is required to avoid spurious matches to e.g. ZCWPW2, which overlaps 19% of ZCWPW1, by requiring >2-fold this match length; although several ZCWPW2 copies are in our initial list, no gene annotated as most similar to ZCWPW2 attains the 39% overlap. For proteins within the "clear" set, we identified the longest alignment for each given species, resulting in 136 species with a likely ZCWPW1 orthologue. This initial screen identified three groups of fish, placental mammals, marsupials, monotremes and reptiles as likely possessing ZCWPW1, similar to our final conclusions. Moreover, it defined amino acids 247-428 in the human reference sequence as being conserved in the alignment (at most 1 sequence not aligning), including the annotated ZCW and PWWP domains.

Using this initial set, we identified additional orthologous sequences and refined our results, by identifying conserved bases within ZCWPW1. Specifically, we divided the 136 species into major clades, as in (Z. Baker et al., 2017), and gave each sequence a weighting so that the overall weight for each clade was the same (so e.g. each Placental mammal sequence was downweighted as this clade was over-represented). Within the consistently aligned region 247-428, we calculated the overall weighted probability of each of the 20 amino acids, or a gap (adding 10^{-5} to exclude zero weights). This identified 31 completely conserved amino acids, and 78 amino acids whose entropy was below 1 (equivalent to 2 amino acids, having equal probability, so implying one amino acid present >50% of the time). Finally, we defined ZCWPW1 orthologues as those sequences matching at least 90% of those 31 perfectly conserved bases which were aligned, and aligning to at least 50% of these bases (this last condition allows for inclusion of incomplete sequencing or protein assembly). We note that while this 90% condition is arbitrary,

by definition all annotated orthologues exceed this threshold, while thresholds below 70-80% are exceeded by orthologues of ZCWPW2 among other genes, offering some justification. Nonetheless, it is worth pointing out that our analysis does not rule out the existence of more poorly conserved ZCWPW1 copies in more distantly related species. This approach identified a final set of 167 genes, which we annotated as likely ZCWPW1 orthologues and used for the majority of results. For each ZCWPW1 orthologue, we also identified the taxonomic relationship of the closest species possessing such a PRDM9 copy . We also reciprocally annotated each PRDM9 copy previously identified (Z. Baker et al., 2017) according to the closest species also possessing ZCWPW1.

Our analyses revealed a relationship between ZCWPW1 predicted functional domains and PRDM9 histone modifications. Therefore, we identified additional potential conserved sites by identifying perfectly conserved bases among those species possessing both ZCWPW1 and PRDM9 orthologues, and where three key SET domain catalytic amino acids within PRDM9 are intact, meaning PRDM9 is predicted to have normal histone modification activity (Z. Baker et al., 2017). This identified a slightly larger number of conserved amino acids (37). Only eight of these varied in any of the 167 species with a potential ZCWPW1 orthologue: 260, 404 and 411 (in two Xenopus frogs, with 411 also in white-headed capuchin), 19 and 22 (in three canines), 25 (in Ocelot gecko), 257 (in Anolis lizard and in Wombat), 325 (in elephantfish). While these changes might alter ZCWPW1 function, the true behaviour of ZCWPW1 is uncertain in these cases, though it is interesting that canids and frogs represent clades that all appear to have lost PRDM9, and possess multiple, clustered, amino acid changes.

3.8.2 Domain Search

pDomThreader (Lobley et al., 2009), was used via the PSIPRED server (`http://bioinf.cs.ucl.ac.uk/psipred/`) on the following uniProt amino acid sequences (Q9H0M4|ZCPW1_HUMAN, Q6IR42|ZCPW1_MOUSE, E2RFJ2|E2RFJ2_CANLF, M3XJ39_LATCH, A0A3S5ZP38_BOVIN, M3WDY6_FELCA, and G3ULT5|G3ULT5_LOXAF)

all of which identified a match to 1ub1A00 in the C terminal section after the PWWP domain, except in LOXAF in which this match was slightly below the p-value threshold of 0.001.

SCP1 prediction used NCBI conserved domain search server (`https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi?`) with the uniProt amino acid sequences Q9H0M4|ZCPW1_HUMAN and Q6IR42|ZCPW1_MOUSE.

3.8.3 Mice and genotyping

KO mice for *Zcwpw1* (C57BL/6N-Zcwpw1$^{em1(IMPC)Tcp}$) were generated by the Toronto Centre for Phenogenomics (Canada). A 1485-bp deletion on chromosome 5 spanning exons 5 to 7 of *Zcwpw1* (chr5:137799545-13780101029) was engineered by CRISPR/Cas9 using guide RNAs 5'-GACTGCACTCACGGCCATCT-3'. The frameshift deletion introduces a stop codon in Exon 8, leading to a predicted unstable short truncated 492bp transcript. Mice were genotyped at the *Zcwpw1* locus using the following primers, and standard cycling conditions: KO allele-Forward, 5'-CACAGGCTCATGTATGTTTGTCTC-3'; KO allele-Reverse, 5'-CTGCTTCGTCCTCTTTCCTTATCTC-3'; WT allele-Forward, 5'-TGCCACCACACTTCATTTG
3'; WT allele-Reverse CCTGTTTCCTTCCCAACTCA-3'. The deletion was verified by direct Sanger sequencing of the KO genotyping PCR product (Source Bioscience, UK), following purification with the QIAquick PCR Purification Kit (Qiagen). Sequence analysis was carried out using Chromas LITE (version 2.1.1).

KO mice for *Prdm9* were described previously (Hayashi et al., 2005) and obtained from the RIKEN BioResource Research Center in Japan (strain name B6.129P2-Prdm9^{tm1Ymat}, strain number RBRC05145). Mice were genotyped at the *Prdm9* locus using the following primers, and standard cycling conditions: WT allele-Forward 5'-AGGAATCTTCCTTCCTTGCTGTCG-3'; WT allele-Reverse 5'-ATTTCCCTGTATCTTCTTCAGGACT-3'; KO allele-Reverse 5'-CGCCATTCAGGCTGCGCAAC
3'.

All animal experiments received local ethical review approval from the University of Oxford Animal Welfare and Ethical Review Body (Clinical Medicine board) and

were carried out in accordance with the UK Home Office Animals (Scientific Procedures) Act 1986.

3.8.4 Fertility measurements

Fertility was assessed in mice ranging from 9 to 12 weeks of age, either by mating with WT littermates and recording the average litter size and frequency, or by measuring paired testes weight (normalized to lean body weight), sperm count (per paired epididymides) and chromosome synapsis rate (by immunostaining of pachytene spermatocytes) in males. Lean body weight was measured using the EchoMRI-100 Small Animal Body Composition Analyzer.

3.8.5 Immunostaining of spermatocytes

Mouse testis chromosome spreads were prepared using surface spreading (Barchi et al., 2008; Peters et al., 1997) and immunostained as previously described (B. Davies et al., 2016). The following primary antibodies were used: custom ZCWPW1 rabbit antiserum (1:100), mouse anti-SYCP3 (Santa Cruz Biotechnology sc-74569, D-1) or biotinylated rabbit anti-SYCP3 (Novus NB300-232), rabbit anti-DMC1 (Santa Cruz Biotechnology sc-74569, D-1), rabbit anti-HORMAD2 (Santa Cruz Biotechnology sc-82192), mouse anti-RAD51 (Abcam ab88572), rabbit anti-RPA2 (Abcam ab10359), mouse (Millipore 05-636, clone JBW301) or chicken (Orbit orb195374, discontinued) anti-phospho γ-H2AX and Alexa Fluor 488-, 647- or 594-conjugated secondary antibodies against rabbit, mouse or chicken IgG (ThermoFisher Scientific), as well as avidin 647 (ThermoFisher Scientific) were used to detect the primary antibodies. Images were acquired using either a BX-51 upright wide-field microscope equipped with a JAI CVM4 B&W fluorescence CCD camera and operated by the Leica Cytovision Genus software, or a Leica DM6B microscope for epifluorescence, equipped with a DFC 9000Gt B&W fluorescence CCD camera, and operated via the Leica LASX software. Image analysis was carried out using Fiji (ImageJ-win64). The specificity of the ZCWPW1 signal was verified using the pre-immune serum from the same rabbit: no staining was visible at any stage of prophase I (data not shown).

3.8.6 FISH

Following immunostaining of spermatocytes, telomeres were labelled using the Telomere PNA FISH Kit/Cy3 (Agilent), following the manufacturer's instructions, but without protease treatment. For the identification of chromosome 18, BAC RP24-144B10 (CHORI, BACPAC Resource Center) was labelled with the Abbot Molecular Nick Translation kit, and applied to the immunostained spermatocytes. The hybridisation and signal detection were carried out using standard techniques. Plasmids

Constructs encoding full-length human (h) or chimp (c) PRDM9 with a C-terminal V5 or N-terminal YFP tag for expression in mammalian cells (pLENTI CMV/TO Puro DEST backbone vector, Addgene plasmid # 17293; Campeau et al., 2009 were described previously (Altemose et al., 2017). Constructs encoding full-length human (h) and mouse (m) ZCWPW1 for expression in mammalian cells were purchased from Genescript (hZCWPW1 cDNA with a C-terminal HA tag in pCDNA3.1; clone ID OHu16813) and Origene (mZCWPW1 cDNA with a FLAG-Myc dual tag in pCMV6-Entry; clone ID MR209594), respectively. For expression in E. Coli, full-length mZCWPW1 cDNA was subcloned into the pET22b(+) vector (Novagen) with a C-terminal poly-His tag.

3.8.7 ZCWPW1 antibody production

pET22b-mZCWPW1-His construct was transformed into BL21 (DE3) E.Coli cells (Invitrogen). Bacterial cultures were grown to a density with O.D600~0.7, and expression of recombinant His-tagged ZCWPW1 protein was induced overnight at 20°C by addition of IPTG to 0.5mM. The protein was purified using Talon metal affinity resin, according to the manufacturer's instructions (Takara). Further purification was carried out by size exclusion chromatography (Superdex HiLoad 200 16/60, GE Life Sciences). Western Blot validation was carried out as described previously (Altemose et al., 2017), using mouse anti-human ZCWPW1 (SIGMA SAB1409478) and mouse anti-polyhistidine (SIGMA H1029). The purified protein was used to immunize 2 rabbits (Eurogentec, Belgium), and the resulting immune antisera were tested against the recombinant antigen by ELISA (Eurogentec), and

the endogenous protein in mouse testes from WT and KO mice (data not shown), alongside the preimmune sera.

3.8.8 Cell line, transfection, and IF staining

Human embryonic kidney (HEK) 293T cells were cultured and transfected using Fugene HD as previously described (Altemose et al., 2017). High and comparable expression of the target proteins was verified by immunofluorescence staining of duplicate transfected cultures before proceeding to ChIP. PRDM9 expression was visualized either directly from live cells through YFP fluorescence, or by immunofluorescence staining of fixed cells using rabbit anti-V5 (Abcam ab9116, 1:500) as described previously (Altemose et al., 2017). ZCWPW1-HA was detected by immunostaining as above, using rabbit anti-HA (Abcam ab9110, 1:100). The fraction of cells co-expressing ZCWPW1-HA and PRDM9-V5 was determined by co-staining with both antibodies, as above (figure A.6).

3.8.9 Immunoprecipitation and Western Blot detection

Cell and testes protein extracts were prepared in lysis buffer containing 50mM Tris-HCl pH8.0, 150mM NaCl, 1% Triton X-100 and a cocktail of protease inhibitors (SIGMA), followed by gentle rotation at 4°C for 30 to 45 min, respectively. Cell debris were pelleted by centrifugation at 4°C for 20 min at 20,000g. Protein extracts were quantified using the BCA assay (Pierce) and equal quantities were incubated overnight at 4°C with 5µl of rabbit polyclonal ZCWPW1 antiserum. Immunocomplexes were pulled down with 50µl of sheep anti-rabbit Dynabeads M-280 (Life Technologies) for 2h at 4°C with gentle rotation. After 5 washes in lysis buffer they were eluted from the beads by boiling in Laemmli sample buffer for 5min, and resolved on either 8% or 4-20% Novex Tris-Glycine precast gels (Invitrogen). Proteins were transferred onto PVDF membranes and the proteins of interest were detected by western blotting following standard procedures. Blots were blocked for 1h at room temperature (RT) in PBS containing 0.2% Tween-20 (PBS-T) and 5% milk, and incubated for 1h at RT with rabbit anti-ZCWPW1 (1:1,000 dilution),

washed 3 washes in PBS-Tween buffer and incubated for 1h at RT with HRP-conjugated sheep anti-rabbit IgG antibody (GE Healthcare). Protein signals were revealed using the ECL Prime western blotting detection reagent according to the manufacturer's recommendations (GE Healthcare). The specificity of the ZCWPW1 signal observed in B6 testis extract was verified by immunoprecipitation using the pre-immune serum from the same rabbit: no signal was detected (data not shown).

3.8.10 ChIP-seq

DMC1 ChIP-seq data from WT B6 mouse testis were generated previously (Brick et al., 2012). H3K4me3 ChIP-seq data was generated previously (B. Davies et al., 2016).

ChIP

ChIP against ZCWPW1-HA, (h/cPRDM9-V5) was carried out from transfected HEK293T cells as previously described (B. Davies et al., 2016; Altemose et al., 2017) were crosslinked for 10min in 1% formaldehyde, the reaction was quenched for 5min by the addition of glycine to a final concentration of 125mM, and the cells were washed twice in cold PBS. The cell pellet was resuspended in cold sonication buffer (50mM Tris-HCl pH8, 10mM EDTA, 1% SDS) supplemented with protease inhibitors (Complete, Roche), and chromatin was sheared to an average size of 200-500bp by sonication for 35 cycles (30s ON/30s OFF) using a Bioruptor Twin (Diagenode). After centrifugation for 10min at 20,000g, 4°C, the sonicate was pre-cleared for 2h at 4°C with 65µl of Dynabeads™ M-280 sheep anti-rabbit IgG (Life Technologies) and a 1% input chromatin sample was set aside. The rest of the sample was diluted 10 fold in ChIP buffer (16.7mM Tris pH8 1.2mM EDTA, 167mM NaCl, 1.1% Triton X-100) supplemented with protease inhibitors, and incubated overnight at 4°C with 5µl of either Abcam rabbit antibody: anti-HA (ab9110), anti-H3K4me3 (ab8580), anti-H3K36me3 (ab9050). Immunocomplexes were washed once with each of low salt buffer (20 mM Tris pH8, 150 mM NaCl, 1% Triton X-100, 0.1% SDS, 2mM EDTA), high salt buffer (20 mM Tris pH8, 500 mM NaCl,

1% Triton X-100, 0.1% SDS, 2mM EDTA), LiCl buffer (10mM Tris pH8, 0.25M LiCl, 1% NP-40, 1% sodium deoxycholate, 1mM EDTA) and TE buffer (10mM Tris pH8, 1mM EDTA) and eluted from the beads in 100 mM NaHCO3, 1% SDS for 30 min at 65°C with shaking. Both input and ChIP samples were reverse crosslinked overnight at 65°C in the presence of 200mM NaCl, and proteins were digested for 90 min at 45°C by addition of proteinase K (0.3ug/ul final concentration). DNA was purified using the MinElute reaction cleanup kit (QIAGEN), and quantified using the dsDNA HS assay kit and a Qubit® 2.0 Fluorometer (Invitrogen).

ChIP against DMC1 was performed from *Zcwpw1*$^{-/-}$ testes using the published method by (Khil et al., 2012) with some modifications listed here. Chromatin shearing was carried out in 20mM Tris-HCl pH8, 2mM EDTA, 0.1% SDS using a Bioruptor Pico sonicator (Diagenode) for 4 cycles of 15s ON/45s OFF. ChIP was performed in 10mM Tris-HCl pH8, 1mM EDTA, 0.1% Sodium Deoxycholate, 1% Triton X-100, 500mM NaCl using 5µg of mouse anti-DMC1 2H12/4 (Novus NB100-2617) pre-bound to 50µl of Dynabeads™ M-280 sheep anti-mouse IgG (Life Technologies) .

Sequencing

ZCWPW1 ChIP and input libraries from transfected cells were prepared by the Oxford Genomics Centre at the Wellcome Centre for Human Genetics (Oxford, UK) using the Nextera DNA Library Prep Kit and established Illumina protocols, and sequenced on a HiSeq 4000 platform (75bp paired end reads, 48 million reads/sample). DMC1 ChIP libraries from *Zcwpw1*$^{-/-}$ testes were prepared and sequenced as described previously (B. Davies et al., 2016) on an Illumina HiSeq2500 platform (Rapid Run, 51bp paired end reads, 110 million reads/sample).

Read Mapping

For the HEK293 experiments, reads were mapped to hg38 (NCBI's GCA 000001405.15 GRCh38 no alt's plus hs38d1 analysis set) using bwa mem (version 0.7.17-r1188) (Li, 2013). Duplicates were removed using picard's markDuplicates (version 2.20.4-SNAPSHOT). Unmapped, mate unmapped, non primary alignment, failing platform,

and low MAPQ reads were removed using samtools with parameters "-q 30 -F 3852 -f 2 -u" (version 1.9 (using htslib 1.9)). Other unmapped reads and secondary alignments were removed using samtools fixmate. Fragment position bed files were created using bedtools bamtobed (v2.28.0).

For the DMC1 mouse experimental data, we processed the data following the algorithm provided by (Khil et al., 2012) to map the reads to the mouse mm10 reference genome (Lunter et al., 2011), and obtain type I reads.

Peak Calling

We called DMC1 peaks, as described previously (B. Davies et al., 2016). For the HEK293 experiments, peaks for ZCWPW1, PRDM9, H3K4me3, and H3k36me3 were called using a peak caller previously described (B. Davies et al., 2016) and available at `https://github.com/MyersGroup/PeakCaller`. Single base peaks were called with parameters pthresh 10^{-6} and peakminsep 250 on 22 autosomes. For each experiment, we used available IP replicates, and sequenced input DNA to estimate background, in calling peaks. Most of our analyses refer to these peaks, except that figures 3.5A plots the change in ZCWPW1 occupancy following PRDM9 transfection, for either the chimp or human PRDM9 alleles. To estimate ZCWPW1 levels with vs. without PRDM9, we called ZCWPW1 peaks in HEK293T cells also transfected with PRDM9 as before, but now replacing the "input" lane with the IP data for ZCWPW1 in cells not transfected with PRDM9. Figure A.8 also uses this measure, of ZCWPW1 recruitment attributable to PRDM9 binding, to predict DMC1 levels, while A.16 (right panel) checks that changes in ZCWPW1 occupancy occur largely at PRDM9-bound sites. We analysed previously published H3K4me3 and SPO11 read data (B. Davies et al., 2016; Lange et al., 2016).

Enrichment Profiles for HEK293T experiments

Peaks were filtered if: center within 2.5kb of PRDM9 independent H3K4me3 (promoters), input coverage <=5, in top 5 by likelihood, greater than 99.9th percentile input coverage. For human PRDM9 peaks the PRDM9 motif position (Altemose et al., 2017) was inferred using the getmotifs function in MotifFinder

within the 300bp region around the peak (with parameters alpha=0.2, maxits=10, seed=42, stranded_prior=T), and peaks were recentered and stranded at these locations. Peaks within 4kb of one another are removed to avoid double counting. Mean coverage was calculated with bwtool v1.0 using the aggregate command with parameters "-fill=0 -firstbase" at a width of ±2kb. Each profile was normalised by the total coverage of all of the fragments as well as by the input profile. Random profiles were created using bedtools random with 10,000 locations, with seed 72346, and were also filtered such that non were within 4kb of one another.

DSB Profiles

Mapped type 1 reads were filtered by removing non-canonical chromosomes (with underscores in names), sex chromosomes, and mitochondrial reads. Bedgraphs were converted to bigWig format using UCSC bedGraphToBigWig v4. Profiles were created using bwtool v1.0 using the aggregate command with parameters 5000:2000 and -fill=0. B6 wild type hotspot locations were filtered to remove non-autosomal chromosomes, retain only B6 allele hotspots, and remove hotspots with PRDM9-independent H3K4me3. *Prdm9*$^{-/-}$ hotspots were filtered to remove X chromosome. Motif centered coordinates were used for wild type hotspots. To normalise background signal, the mean signal between -5000 and -3000 was subtracted from each strand-mouse combination. Mean coverage was normalised by the sum of coverage over each strand-mouse combination across both WT and KO locations.

Mapping of Alu CpGs

Alu locations were downloaded from UCSC tables and filtered for Alu repeats with a width between 250 and 350bp. DNA sequence was extracted from the genome (Ensembl 95 h38 primary assembly) at these locations using the bedtools getfasta command and CpG dinucleotides were counted using 'stringr::str_locate_all' command in R.

CpG Methylation

Data is from (Libertini et al., 2015) (GSE51867), non CpG methylated sites were removed, and the two strands were summed. High copy number (>1.5) CpGs were removed.

DMC1 Prediction using ZCWPW1 and PRDM9 binding strength

PRDM9-dependent ZCWPW1 was force-called at the human PRDM9 peaks. DMC1 sites are from (Pratto et al., 2014b) (GSE59836) and were subsetted to 'AorB' autosomal peaks with a width less than 3kb. These peaks were then trimmed to a maximum of 800bp before overlapping with PRDM9 peaks to create a binary target variable. Regions with input coverage <=5 in either predictor were removed in addition to outliers with input > 200 and or Zcwpw1 enrichment >10. Chromosomes 1, 3, and 5 were used as the test data with the remaining autosomes being the training data.

Heatmaps

Regions were extracted from bigWig using bwtool matrix -fill=0 -decimals=1 -tiled-averages=5, and width parameters 2000. Coverage was normalised by total coverage (scaled by 10^{10}). A pseudocount of 1 was added to both input and sample, and the sample was then normalised by the input for each region. Values outside the quantile range 0.01-0.99 were thresholded. The profile plots are created by taking the ratio of the mean coverage of the sample and input separately. Ordering of the regions was determined by the mean coverage of a 200bp window centered on the peak center.

Software

Computational analysis was performed using R (R Core Team, 2018), and Snakemake (Köster et al., 2012). BigWig aggregations by bwtool (Pohl et al., 2014), BED manipulation with BEDtools (Quinlan et al., 2010), and FASTA processing with seqkit (W. Shen et al., 2016). We also used BigWig tools (Kent et al., 2010), IGV (Thorvaldsdóttir et al., 2013), and GNU parallel (Tange, 2011). Plots were created

using the ggplot2 package (Wickham, 2016) and extensions ggforce (Pedersen, 2016), RColorBrewer (Neuwirth, 2014), viridis (Garnier, 2018), drawProteins (Brennan, 2018), and cowplot (Wilke, 2018). In addition the following R packages were used: data.table (Dowle et al., 2019), mgcv (S. N. Wood, 2011), PRROC (Grau et al., 2015), sangerseqR (Hill et al., 2014), and ComplexHeatmap (Z. Gu et al., 2016).

3.8.11 Data and Code availability

Code is available at `www.github.com/MyersGroup/Zcwpw1` archived at Zenodo (DOI: 10.5281/zenodo.3559758). Raw and processed data from ZCWPW1 and DMC1 ChIPseq is available at GEO accession GSE141516.

DMC1 SSDS ChIPseq from human testis is from (Pratto et al., 2014b) (GSE59836). H3K4me3 ChIPseq from mice testis is from (B. Davies et al., 2016) (GSE73833). SPO11 ChIPseq from mice testis is from (Lange et al., 2016) (GSE84689). DMC1 SSDS ChIPseq from WT B6 mouse testis is from (Brick et al., 2012) (GSE35498). PRDM9, H3K4me3 and H3K36me3 ChIPseq from HEK293 cells is from (Altemose et al., 2017) (GSE99407). Whole-genome bisulfite sequencing (BS-seq) of HEK293 cells is from (Libertini et al., 2015) (GSE51867).

S'il se produit la quatre spores et pas deux, c'est qu'elles doivent avoir chacune quelque chose de particulier.

If four spores occur and not two, then each must have something special.

— Janssens, 1909

4 Perspectives

Contents

4.1 Other Single Cell RNAseq studies 133
4.2 Observational Extensions 139
4.3 Other Zcwpw1 Papers & Extensions 140

4.1 Other Single Cell RNAseq studies

Whilst there were no single cell RNAseq studies of spermatogenesis when this research was initiated, it was clearly a popular idea as there are now an increasing number of such studies (22 as of December 2019), briefly reviewed and compared with our results below. Of note, only two other studies examine mutant phenotypes, and one other study examines transcription factor binding: we discuss these further below. Finally no other study has used an SDA based approach.

Building on a previous study (J. Guo et al., 2017) of scRNAseq of 92 cells from human SSEA4+ hSSCs and c-KIT+ spermatogonia, J. Guo et al., 2018 generated a dataset of 6,490 cells using the 10X Chromium platform from 3 human donors. This study performed RNA velocity analysis on early germ cells, which revealed two sub-populations of the earliest stem cells (one steady, one committing).

Study	Cells	Organism	Method
Hermann et al., 2018	62,141	Human & Mouse	Chromium & SMARTer (C1)
Ernst et al., 2019	53,510	Mouse	Chromium
Green et al., 2018	34,633	Mouse	Drop-seq
Sohni et al., 2019	33,585	Human	Chromium
Jung et al., 2019	20,322	Mouse	DropSeq
Han et al., 2018	19,659	Mouse	Microwell-seq
Grive et al., 2019	15,882	Mouse	Chromium
Law et al., 2019	10,140	Mouse	Chromium
La et al., 2018	9,424	Mouse	Chromium
Fang et al., 2019	6,804	Mouse	Chromium
J. Guo et al., 2018	6,490	Human	Chromium
B. Xia et al., 2019	4,147	Human & Mouse	inDrop
M. Wang et al., 2018	3,028	Human	Smart-seq2
Lukassen et al., 2018	2,550	Mouse	Chromium
Vértesy et al., 2019	1,274	Mouse	SORT-seq
Y. Chen et al., 2018	1,174	Mouse	Smart-seq2
Stévant et al., 2018	400	Mouse	SMARTer (C1)
Song et al., 2016	201	Mouse	SMARTer (C1)
Makino et al., 2019	175	Mouse	SMARTer (C1)
Neuhaus et al., 2017	105	Human	Tang
J. Guo et al., 2017	92	Human	SMARTer (C1)
Liao et al., 2019	71	Mouse	SMARTer (C1)
Total	285,807		

Interestingly, it also revealed later stem cells whose velocity vectors were pointing "backwards" towards less differentiated state (J. Guo et al., 2018), consistent with previous reports of plasticity in spermatogonia (Brawley et al., 2004; Nakagawa et al., 2010; Hara et al., 2014).

They highlight that *Csf1r* was reported to be expressed only in spermatogonia in mice, but in their human data it is specifically expressed in macrophages (J. Guo et al., 2018). However, this is not what the original study claimed, rather that there exists a very rare population of THY1+ spermatogonial stem cells that also express low levels of *Csf1r*, in addition to the macrophage expression (Oatley et al., 2009). Indeed, our mouse data shows specific *Csf1r* expression in macrophages, and in our dataset very few cells have detectable Thy1. They also highlight that in their human data *Cxcl12* is detected in Leydig cells, but was previously reported as expressed in Sertoli cells in mice (Q.-E. Yang et al., 2013). However, the cited study

only claims that CXCL12 is expressed in Sertoli cells *within* the adult tubule, and it also shows that CXCL12 is detected outside the tubules in Leydig cells. In addition the expression in Sertoli cells was weakly co-localised, confined to the basement membrane, and shown using a poor marker of Sertoli cells (GATA4).

Stévant et al., 2018, isolated a total of 400 NR5A1-GFP+ cells by FACS from E10.5, E11.5, E12.5, E13.5, and E16.5 testis, revealing that both Sertoli and Leydig cells originate from a single common progenitor population, with Sertoli cell differentiation potentially driven by a pulse of expression of *Sry*, *Kdm3a*, and *Nrob1* among others.

Y. Chen et al., 2018, generated a dataset of 1,136 cells over 20 time points from mice with Vasa-dTomato and Lin28-YFP which were treated with retinoic acid to synchronise spermatogenesis, in addition to 38 cells from *Spo11*$^{-/-}$ mice. In agreement with our results they found enrichment of *Prdm9*, *Spo11*, *Gm960*, *Meiob*, *Dmc1*, and *Mcm8* in their leptotene & zygotene cluster C3. They also detected novel enriched genes in this cluster *Fbxo47*, *Pparg*, and *Ccnb3*. *Fbxo47*$^{-/-}$ mice were completely infertile, similar to the previously studied *C. elegans* homologue *prom-1* (Jantsch et al., 2007). A follow up study found *Fbxo47* to be required for DSB repair and synapsis at telomeres, potentially due to inhibition of TFR2 ubiquitination (Hua et al., 2019). To my knowledge, this represents the only other meiotic study (apart from our own) so far where a gene discovery from single cell data resulted in a new publication.

Y. Chen et al., 2018, also identified Sox30 as highly expressed in pachytene and generated *Sox30*$^{-/-}$ mice, which were infertile, in agreement with an earlier study (Feng et al., 2017), and SOX30 ChIPseq detected some peaks in promoter regions for genes that were differentially regulated in *Sox30*$^{-/-}$ mice. Y. Chen et al., 2018, were also able to profile alternative splicing events, enabling characterisation of *Spo11* isoform expression (β in leptotene and α in pachytene). They also investigated MSCI and identified 150 genes as "escaping" MSCI (defined by them as expression in diplotene stage). However, none of these genes show convincing evidence of MSCI escape in our dataset. Vértesy et al., 2019, also find genes escaping MSCI

in a dataset of 1,274 cells from Dazl-GFP mice (up to the start of pachytene, median transcript count 20,488/cell), particularly *Slitrk2*. However, this gene was below the detection threshold in our dataset.

Green et al., 2018 used DropSeq to generate a dataset of 34,633 cells from mice testis (average 6,205 UMIs/cell). They were able to identify two new somatic populations which they describe as an innate lymphoid type II immune cell, and a mesenchymal cell. Whilst it's unclear if our dataset also contains the immune cell based on the marker genes they identified, it most certainly contains the mesenchymal cell which they validated using a Tcf21-creERT2; tdTomato mouse. We identified these cells as telocytes based on previous work defining the markers of these cells (Marini et al., 2018). By using Sox9-EGFP and Amh-cre;mTmG transgenic lines Green et al., 2018 were able to enrich for Sertoli cells, resulting in 9 sub-clusters. Interestingly they also found Prm2 as a marker of some Sertoli cells, and with intronic/UTR probes were able to determine that these transcripts were transcribed in round spermatids but persist in Sertoli cells after phagocytosis.

Green et al., 2018 is the only other study to perform transcription factor motif analysis in addition to our own, and they reach some different conclusions worth discussing. Compared to our motif analysis, one major difference is that they are using MEME-ChIP (which is a combination of MEME, DREME [a non probabilistic regular expression based motif finder], and CentriMo [which looks for central enrichment of known motifs]). The CentriMo manual cautions that when the number of sequences is large, motifs that are only slightly similar can show significant enrichment. It appears that many of their *unknown* motifs are from the DREME method. Our most commonly found motif is the Sp1 family motif, and while they don't find this they do find a reverse complemented version which they assign to Bcl6b. However, the Bcl6b motif in the HOCOMOCO database does not appear greatly similar to the motif they find. Some motifs which they classify as unknown clearly match motifs we found with known transcription factors such as ZNF143 in their gene group 1. They also find a motif which matches well to our CREM-t motif in their group 6 but they do not link it to CREM, likely

due to the short motifs that DREME typically produces. Many of their other unknown motifs are however very long with limited information score diversity. In fact they are close to the maximum length possible for DREME. It is plausible these "motifs" are actually just the promoter sequence of ampliconic genes such as the tasukan family of genes on chromosome 14 and so their enrichment does not imply specific transcription factor binding.

Grive et al., 2019 sampled a total of 15,882 cells form five different postnatal timepoints during the initial wave of spermatogenesis (PND 6, 14, 18, 25, and 30) and so were able to profile how gene expression changes during this testis maturation process as well as asses the proportions of different cells types at each stage. For example DNA repair genes such as *FancJ* (*Brip1*), *Brca1*, *Rad51*, and *Atm* apparently have increased expression in adult testis compared to the equivalent stage during the first wave of spermatogenesis.

Ernst et al., 2019, generated a dataset of 53,510 single cells from both adult (8-9 weeks) and juvenile (PND 5–35, in 5 day steps) mice. These results agree with our own including for example the identification of *Pou5f2* as a marker of late prophase I. In addition, they profiled mice with an additional chromosome - human chromosome 21 (Tc1 mice) - but detected minimal differences in transcription other than relative lack of post-meiotic cell types due to metaphase arrest, reminiscent of our *Mlh3* mutants. Ernst et al., 2019 also looked at MSCI dynamics and in agreement with our results also found a sharp drop in sex to autosome expression ratio at the zygotene-pachytene transition followed by a gradual reactivation. They also highlighted the Ssxb family as one of the first to be expressed post MSCI, and confirmed expression using ISH. This study also performed CUT&RUN to assay H3K4me3, H3K9me3, and K3K27ac, confirming an enrichment of H3K9me3 on the X chromosome of spermatids, and at spermatid specific genes in meiosis, likely deposited by SETDB1 (Hirota et al., 2018).

B. Xia et al., 2019 generated a dataset of 4,147 cells from both human and mouse testis (average UMI 7,459) and note that >90% of all protein coding genes are expressed in germ cells (c.f. 62% in somatic cells), in line with previous

observations (Soumillon et al., 2013; Schmidt, 1996). They propose a model of pervasive transcriptional scanning in the testis germ cells in order to promote transcription coupled repair and hence reduced germline mutations. Consistently they found reduced mutation rates in expressed vs unexpressed genes within the testis and in addition this effect was higher on the template strand (and on the coding strand upstream). They also found human cells had much higher expression of some genes, such as *CXCL6* and *GAPDH*, whereas mouse had higher expression of *Fabp9* and *Sord* (B. Xia et al., 2019).

Law et al., 2019, generated a dataset of 10,140 cells from E16.5, P0, P3, and P6 ID4-eGFP mice (median 20,546 UMI/cell), and showed that ID4-eGFP+ prospermatogonia from E16.5 mice were able to establish colonies in adult germ cell depleted recipient testis.

Fang et al., 2019, generated a dataset of 3,659 wild type and 3,145 *Akap4*$^{-/-}$ cells, a major component of the fibrous sheath of spermatids (Eddy et al., 2003). With our own, this is the only other study, in addition to the human chromosome 21 knock in by Ernst et al., 2019, that performed single cell RNAseq on mutant mice.

Lukassen et al., 2018, generated a dataset of 2,550 cells from adult mice. They claim high expression of *Pou5f1* in round spermatids but no expression was seen at this stage in our study. They also detect expression of *Kit* in round spermatids, and indeed in our study we detect the occasional transcript at this stage.

M. Wang et al., 2018, sequenced 2,854 cells from normal humans, and 174 from a nonobstructive azoospermia patient. Many of the marker genes highlighted and validated by ISH agree with our assignments in mice including *Hmga1* in spermatogonia, *Ovol2* in pachytene and diplotene, and *Tex29* in early spermatids. Sohni et al., 2019, generated a dataset of 18,723 and 14,862 cells from human adult and newborn testis respectively and so were able to compare and investigate neonatal germ cell development.

In comparison to other datasets our dataset has the lowest UMI count per cell (1,312) - 1/5th of the other DropSeq dataset, and 1/10th of the 10X Chromium datasets. Despite this, by using matrix factorisation rather than hard clustering

as in the other studies, it appears we are able to gain as many and often more biological insights from our data.

4.2 Observational Extensions

In total these studies represent over 275,000 cells and so there may be significant advantages in combining them. One challenge of this would be dealing with batch effects from different species, genetic backgrounds, developmental time points, cell capture technology, and sequencing platforms. We have shown SDA can capture batch effects and so could provide one possible solution to this problem, although we have not tested for example combining 10X and DropSeq or human and mouse. Even within one of these studies there could be significant advantages to using a matrix factorisation approach. For example with multiple developmental time points you might expect to find a main component for each cell type, and then "modifier" components representing how the transcription of that cell type changes before, during and after the first wave of spermatogenesis.

Whilst some studies (Ernst et al., 2019) have used other (bulk) assays alongside scRNAseq, there is much to be gained from single cell resolution profiling of protein abundance, protein-DNA binding, chromatin accessibility, methylation/DNA modifications, nuclear structure, and spatial aspects; even more so when these modes are profiled simultaneously in single cells. The main challenge for this has been experimental limitations but now many methods are available to profile one or multiple of these aspects (reviewed in Chappell et al., 2018; Hu et al., 2018; Stuart et al., 2019; Hériché et al., 2019). SDA is also able to perform group factor analysis (Hore et al., 2015), and a similar analysis package has been applied to single cell multiomic data (Argelaguet et al., 2019).

Chromatin accessibility assays would be particularly interesting given the dramatic chromatin remodelling that occurs during spermatogenesis. The genome also undergoes dramatic de-methylation and re-methylation during gametogenesis and so methylation assays would also be intriguing. Protein-DNA assays such as single cell ChIP-seq could help to directly elucidate the gene regulatory programme

in much more detail than we have been able to achieve here by inferring motifs. Due to the spatial layout of the different stages of spermatogenesis in addition to the cycle of the seminiferous epithelium, spatial sequencing seems particularly apt for the investigation of transcriptome of the testis. The spatial structure would help to more precisely state the cell identity of ambiguous transcriptomes, and would help to reveal additional structure, for example how the supporting Sertoli cells differ at different stages of the seminiferous cycle. Many single cell (and even bulk) studies of the testis have focused on the spermatogonia in an effort to identify the true spermatogonial stem cell. Single cell genetic lineage tracing could help to answer this question by providing a molecular readout of the cell division history for each cell (reviewed in Baron et al., 2019; McKenna et al., 2019). For many of these aspects there is already some information known for example which motifs are present at which promoters, and which genes cause infertility or are involved in meiotic processes. Incorporating this prior and multi-omic information into the decomposition would likely help to disambiguate co-expressed gene sets into finer more distinct functional groupings, in addition to providing insight into the molecular processes that underlie the functional changes that occur during the hugely complex and intricate performance that is spermatogenesis.

4.3 Other Zcwpw1 Papers & Extensions

Prior to our study the only work linking *Zcwpw1* to meiosis was the study by Soh et al., 2015, who identified that *Zcwpw1* was specifically expressed in foetal prophase I in female mice. This led to the work by M. Li et al., 2019 revealing that *Zcwpw1* is required for fertility in males, and this work was compared to our results in chapter 3. Since finishing our study, two other reports have been made public investigating the role of *Zcwpw1* in meiotic recombination.

One of these studies, by Mahgoub et al., 2019, was motivated by a reanalysis of the data from Y. Chen et al., 2018, showing high co-expression of *Zcwpw1* with *Prdm9*. They also show co-evolution of *Zcwpw1* with *Prdm9*, but our work

specifically highlights the association with the SET and SSXRD domains of *Prdm9* (required for methylation) compared to the other domains.

Both studies used the same *Zcwpw1*$^{-/-}$ mouse, and both studies found DSB positioning is unchanged. They differ in that Mahgoub and colleges used ENDseq, showing altered DSB repair post homologue invasion, while we used DMC1 SSDS showing altered repair timing at individual hotspots and its association with PRDM9 binding.

The two studies also investigate different model organisms. We used an *in vitro* system to study the human ZCWPW1 and PRDM9 proteins, identifying >800,000 ZCWPW1 binding sites ($p < 10^{-6}$), the top ~10,000 of which are almost all PRDM9-bound sites. Mahgoub and colleagues used mice, identifying 4,300 ZCWPW1 binding sites ($p < 10^{-3}$), again mainly PRDM9-bound sites. The additional weaker peaks we found revealed a more subtle CpG influence on ZCWPW1 binding, which also impacts stronger peaks. While we used chimp PRDM9 to show allele specificity, Mahgoub and colleagues used hybrid mice.

Some analyses were unique to each study. For example we counted stage specific DMC1 by immunofluorescence, revealing DSB repair delay. We also investigated the subnuclear positioning of ZCWPW1 using immunofluorescence, discovering an interesting new telomeric localisation. Using biotin-streptavidin pulldown assays, Mahgoub and colleagues show that *dual* modified H3K4me3-H3K36me3 peptides had the highest binding affinity for ZCWPW1. While we showed that PRDM9 bound sites (with the dual mark) are stronger recruiters of ZCWPW1 than H3K4me3 alone.

Another group, having previously generated a different *Zcwpw1*$^{-/-}$ mouse (M. Li et al., 2019), generated a new mutant with three point mutations in the zf-CW domain: *Zcwpw1*$^{W247I/E292R/W294P}$ rendering the H3K4me3 recognition capacity non-functional (Huang et al., 2019). This mouse was also infertile with complete testicular azoospermia and incomplete synapsis observed by SYCP1/3 staining. ChIPseq against ZCWPW1 revealed expected high overlap with H3K4me3 and DMC1 marks in WT testis and lack of peaks in *Zcwpw1*$^{-/-}$ and *Zcwpw1*$^{W247I/E292R/W294P}$ mice.

ChIPseq against ZCWPW1 in *Prdm9*$^{-/-}$ mice resulted in very few peaks of which >80% were within 5kb of a TSS.

While these papers clearly agree that ZCWPW1 aids DSB repair, exactly how this is achieved remains unclear. In addition, the missing link between PRDM9 and SPO11 recruitment remains missing. As previously discussed *Zcwpw2* is a promising candidate for this role, just as *Zcwpw1* was, and further work will prove what (if any) involvement *Zcwpw2* has in meiotic recombination.

It is likely that many other genes revealed by single cell RNA sequencing will be key players in meiosis and spermatogenesis and hopefully this work will enable some of the many mysteries to be unveiled.

Appendices

In asexual organisms, before the descendants can acquire a combination of beneficial mutations, these must first have occurred in succession, within the same lines of decent.

— Muller, 1932

A

Supplementary Figures

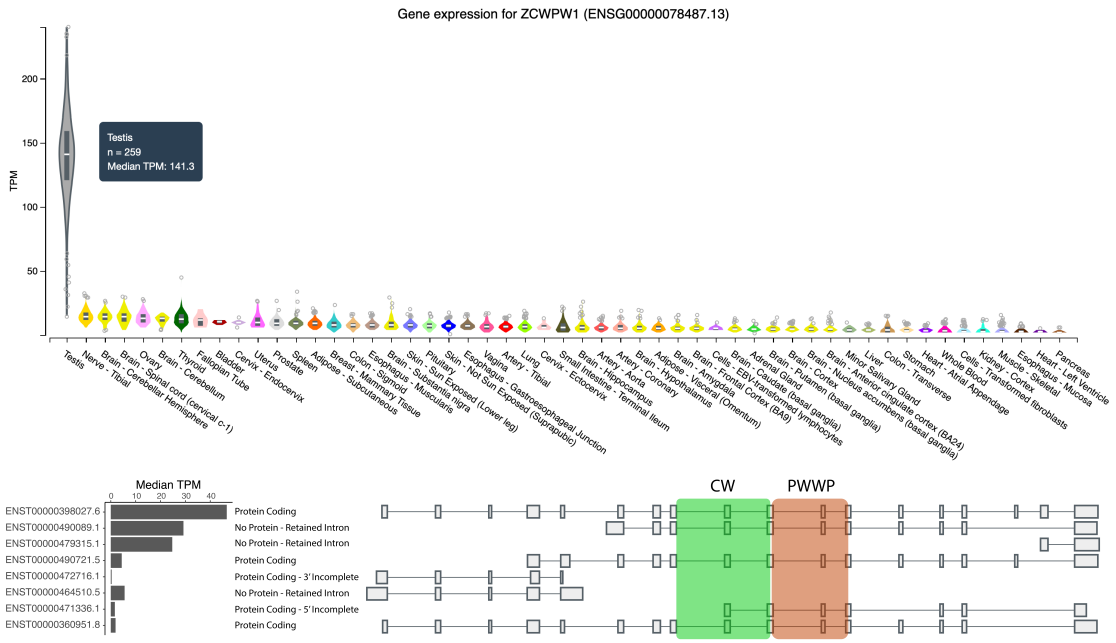


Figure A.1: (A) *Zcwpw1* is highly and specifically expressed in the testis. Data Source: GTEx Analysis Release V7 (dbGaP Accession phs000424.v7.p2). **(B)** *Zcwpw1* isoform expression in human testis.

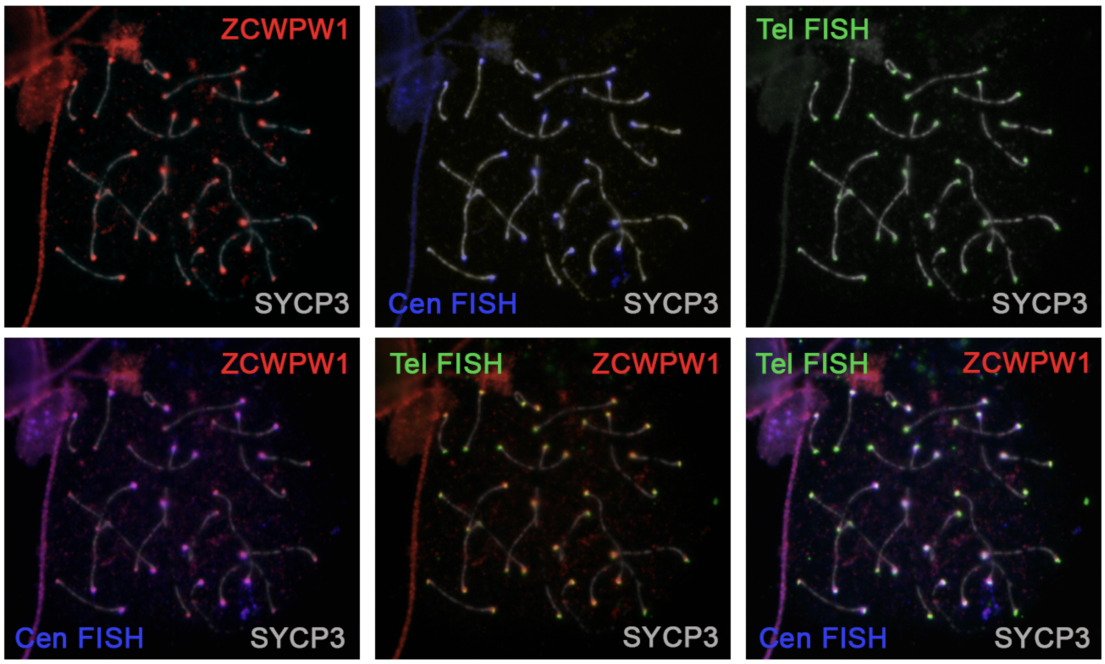


Figure A.2: ZCWPW1 localises at subtelomeric and subcentromeric regions of chromosomes in pseudo-pachytene cells. Testis chromosome spreads from WT mice were immunostained for SYCP3 and ZCWPW1, and hybridised by FISH with a telomeric or centromeric probe. Note that the ZCWPW1 foci do not always exactly co-localize with the telomeric signal, and generally lie more internally on the chromosome axis.

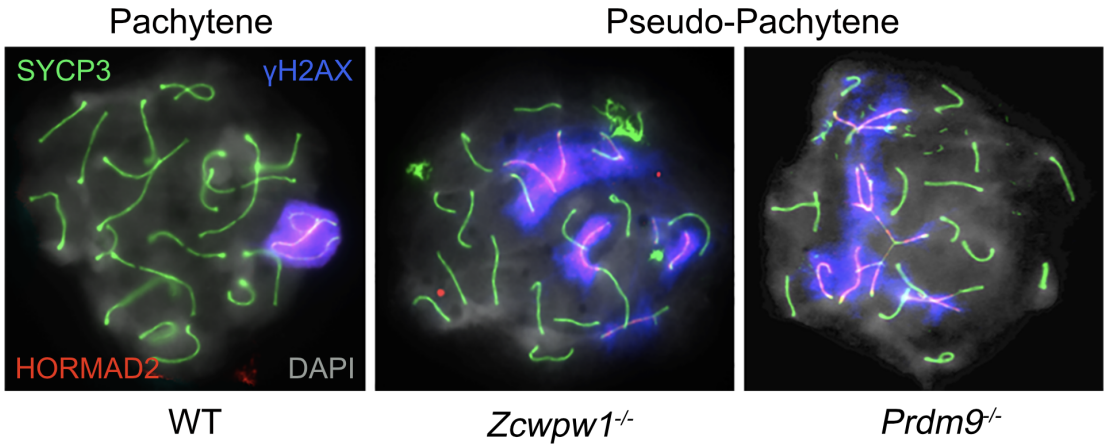


Figure A.3: Asynapsis and lack of XY body formation in *Zcwpw1*$^{-/-}$ mouse testis. Testis chromosome spreads from WT, *Zcwpw1*$^{-/-}$ and *Prdm9*$^{-/-}$ mice were immunostained with antibodies against SYCP3, γ-H2AX (phosphorylated form) and HORMAD2, and counterstained with DAPI. WT Pachytene cells show full synapsis of all autosomes and an XY body strongly labelled with γ-H2AX. In contrast, the XY body is absent in *Zcwpw1*$^{-/-}$ and *Prdm9*$^{-/-}$ pseudo-Pachytene cells, with no clear XY body formation. In the *Prdm9*$^{-/-}$ mutant, mispairing of homologues is evident by the formation of branched structures referred to as "tangled" chromosomes in the text.

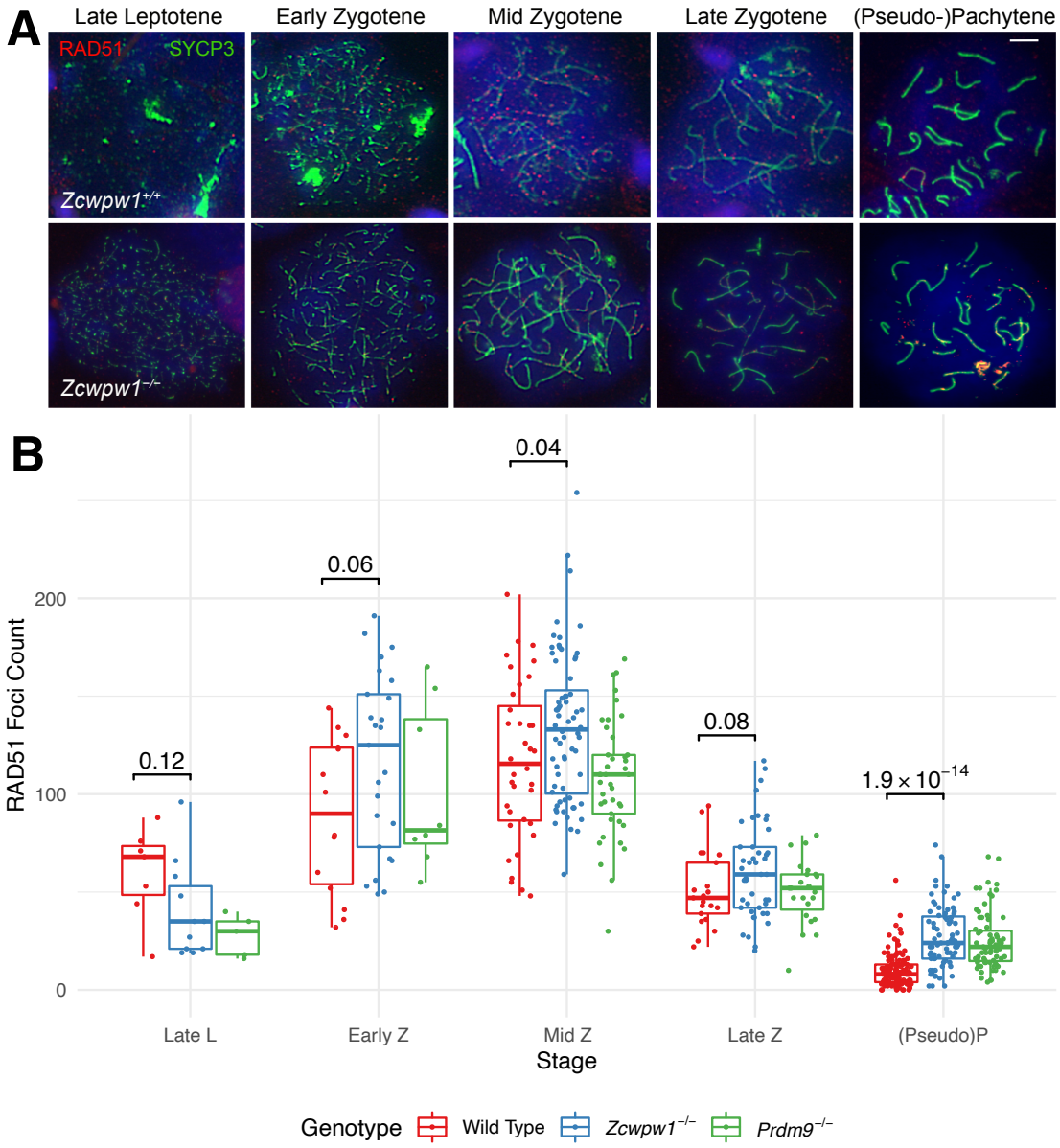


Figure A.4: **(A)** RAD51 staining in *Zcwpw1*$^{-/-}$ mouse testis. Testis chromosome spreads from *Zcwpw1*$^{+/+}$ and *Zcwpw1*$^{-/-}$ mice were immunostained with antibodies against the synaptonemal complex protein SYCP3 and the recombinase RAD51, and counterstained with DAPI to visualise nuclei. Developmental stages are indicated at the top. Scale bar 10 µm. **(B)** The number of RAD51 foci in cells from the various stages of prophase I were counted. p-values are from Welch's two sided, two sample t-test. L: Leptotene, Z: Zygotene, P: Pachytene. n=2 mice per genotype (*Zcwpw1*$^{-/-}$ and WT), n=1 for *Prdm9*$^{-/-}$.

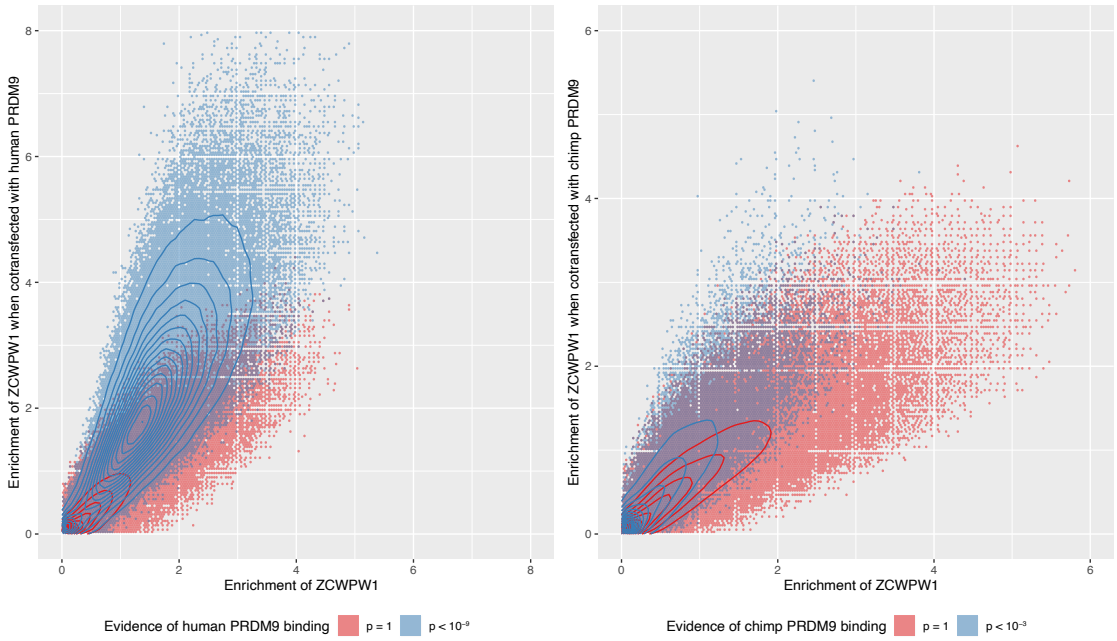


Figure A.5: Enrichment of ZCWPW1 when co-transfected with PRDM9 is dependent on the ability of ZCWPW1 to bind in the absence of PRDM9. Enrichment was force called in 100bp windows across the genome. Data is conditioned on having input coverage of greater than 5 and enrichment>0.01 for both axes. Hexagons are coloured if at least 3 data points are present. Solid lines show density contours estimated by MASS::kde2d() in R.

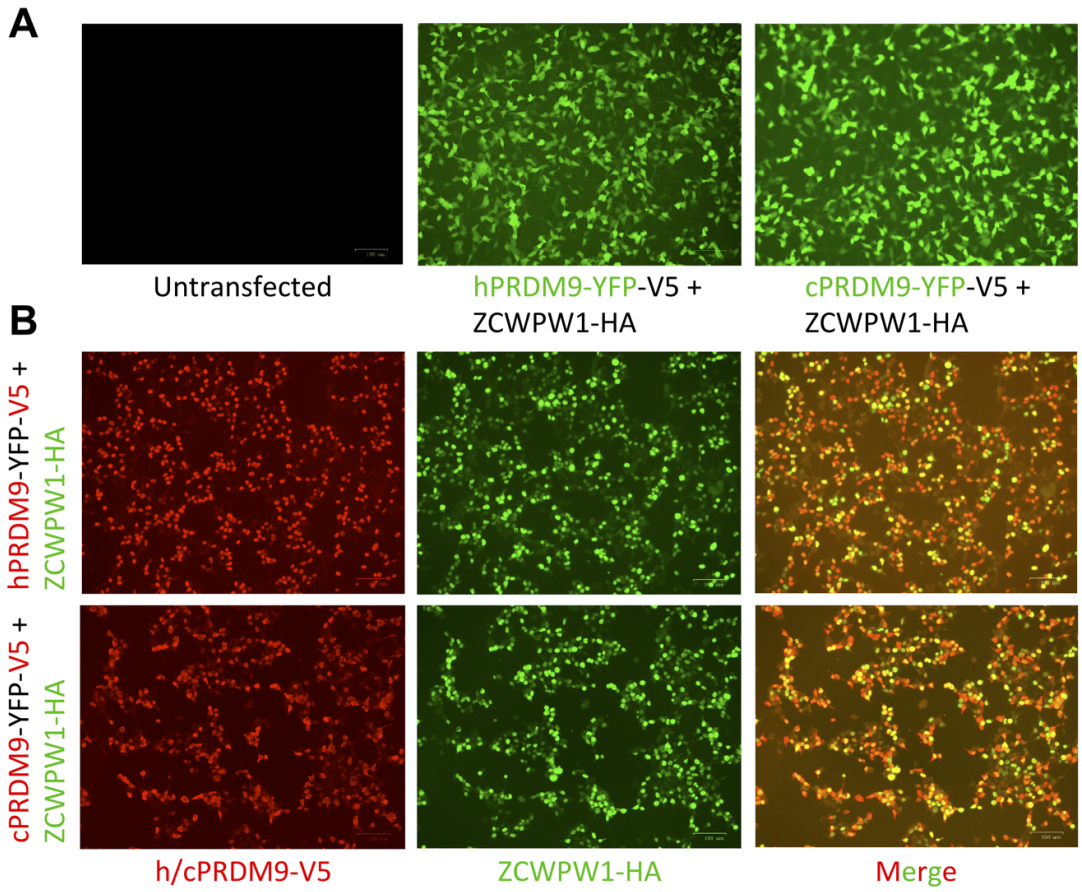


Figure A.6: Co-transfection of ZCWPW1 and PRDM9 in HEK293T cells. Immunofluorescence staining against the protein tags shows high expression levels of each protein and a reasonable proportion of co-expressing cells with merged overlapping signals (ranging from light green to yellow and light red depending on the expression ratio of the two proteins).

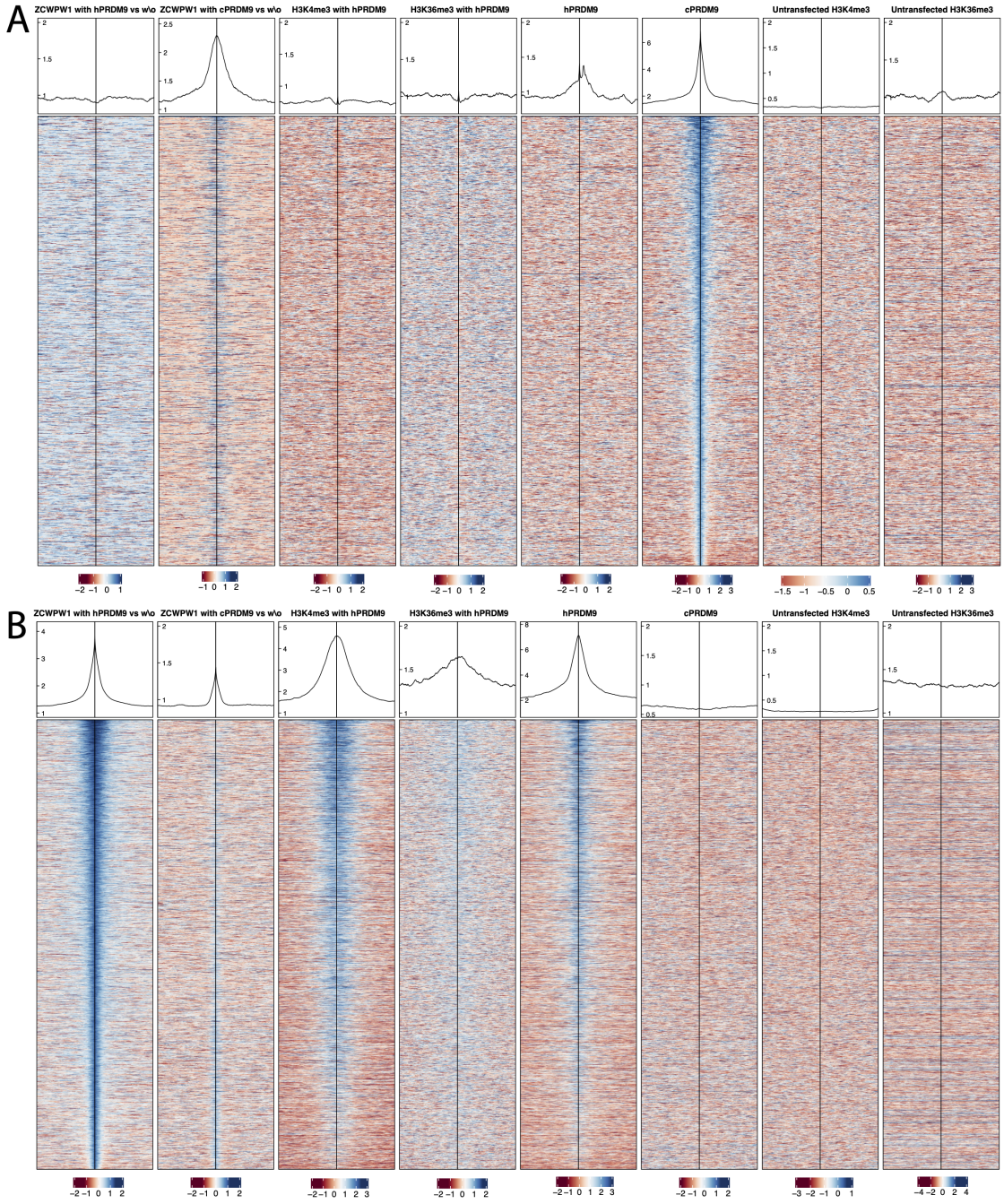


Figure A.7: (A) Profiles and heatmaps of reads at locations of chimp PRDM9. Heatmaps show log fold change of sample (as indicated in the title of each column, methods) vs input, for the top ¼ of peaks of chimp (c) PRDM9, for various samples, ordered by chimp PRDM9. Here showing that ZCWPW1 is found at sites of chimp PRDM9 peaks, when co-transfected with chimp PRDM9, but not at human (h) PRDM9 peaks. **(B)** Profiles and heatmaps of reads at locations of ZCWPW1 co-transfected with human PRDM9. Heatmaps show log fold change of sample (as indicated in the title of each column, Methods) vs input, for the top ¼ of peaks of ZCWPW1 when co-transfected with PRDM9, for various samples, ordered by first column. Here showing that H3K4me3, H3K36me3 and hPRDM9 are found at ZCWPW1 peaks when co-transfected with PRDM9.

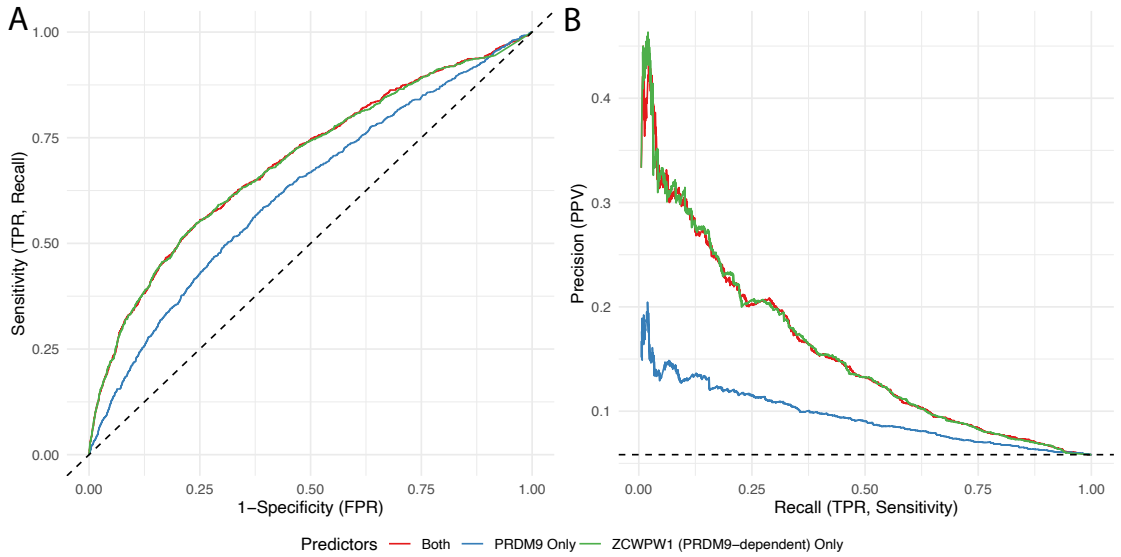


Figure A.8: ZCWPW1 enrichment (with PRDM9 vs without) provides a better predictor of DMC1 sites than PRDM9 itself using a logistic regression (methods section 3.8.10). **(A)** Receiver Operating Characteristic curve **(B)** Precision Recall Curve. ZCWPW1 enrichment (with PRDM9 vs without) refers to enrichment of ZCWPW1 cotransfected with PRDM9 relative to (using as input) ZCWPW1 transfected alone.

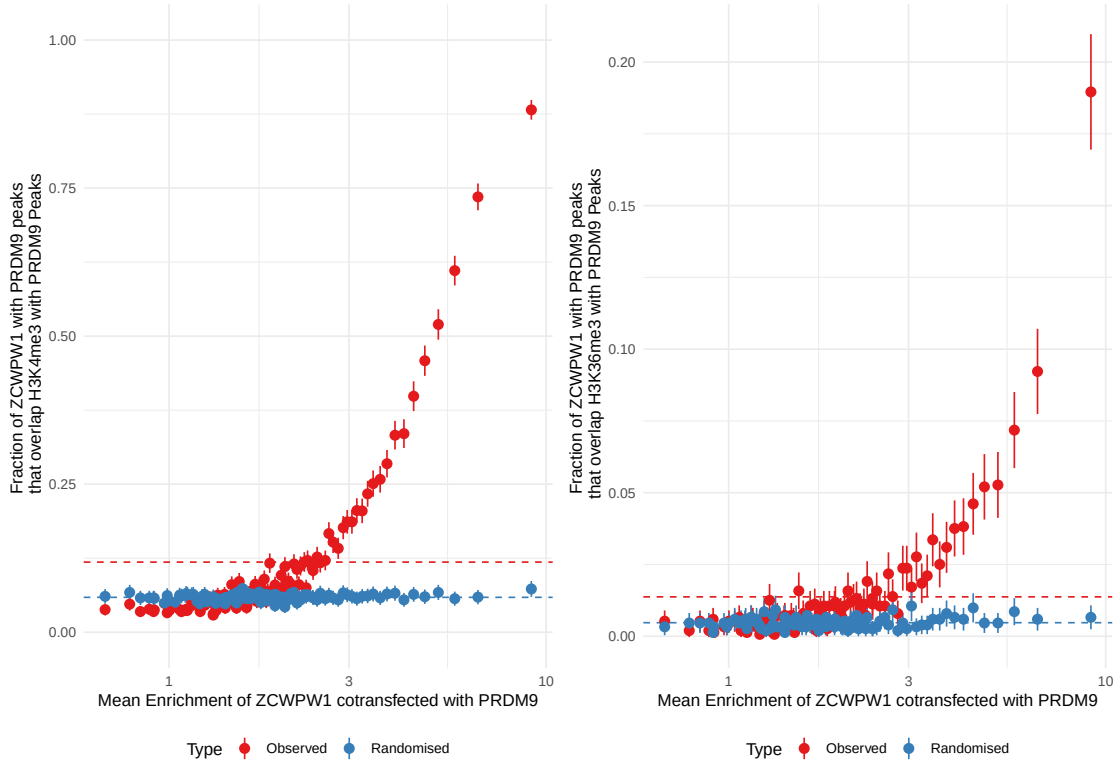


Figure A.9: Fraction of ZCWP1 peaks (co-transfected with PRDM9 with input coverage of at least 5) that overlap either H3K4me3 or H4K36me3 for different bins of ZCWPW1 enrichment (100 equal sample size bins). Error bars show ±2 s.e. of the proportion. "Randomised" shows expected proportions when x-axis regions are randomly shifted within a range of 10^8 bases (up to a maximum of the length of the chromosome).

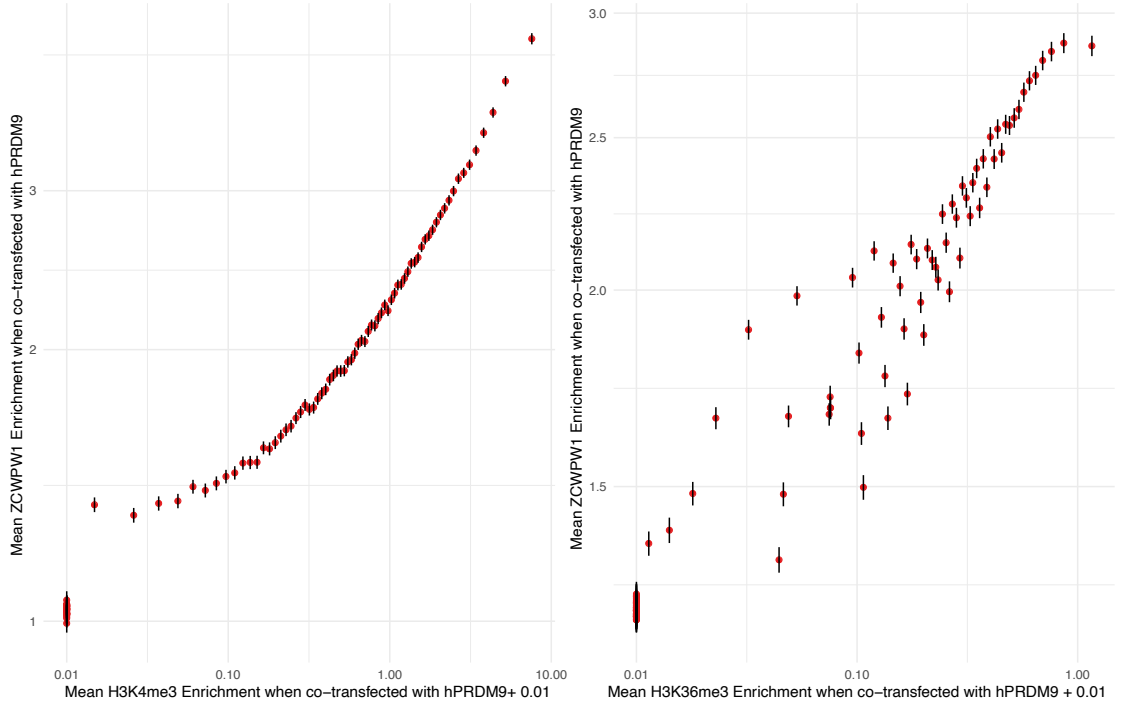


Figure A.10: Enrichment from 100bp non-overlapping windows is binned into 100 equal sample size bins and mean enrichment of ZCWPW1 cotransfected with PRDM9 is plotted. Error bars show ±2 s.e. of the mean. Windows with evidence of PRDM9-independent H3K4me3 have been removed from the H3K4me3 plot. Additionally, x-axis regions were removed if input reads were <15 and y-axis regions if <5.

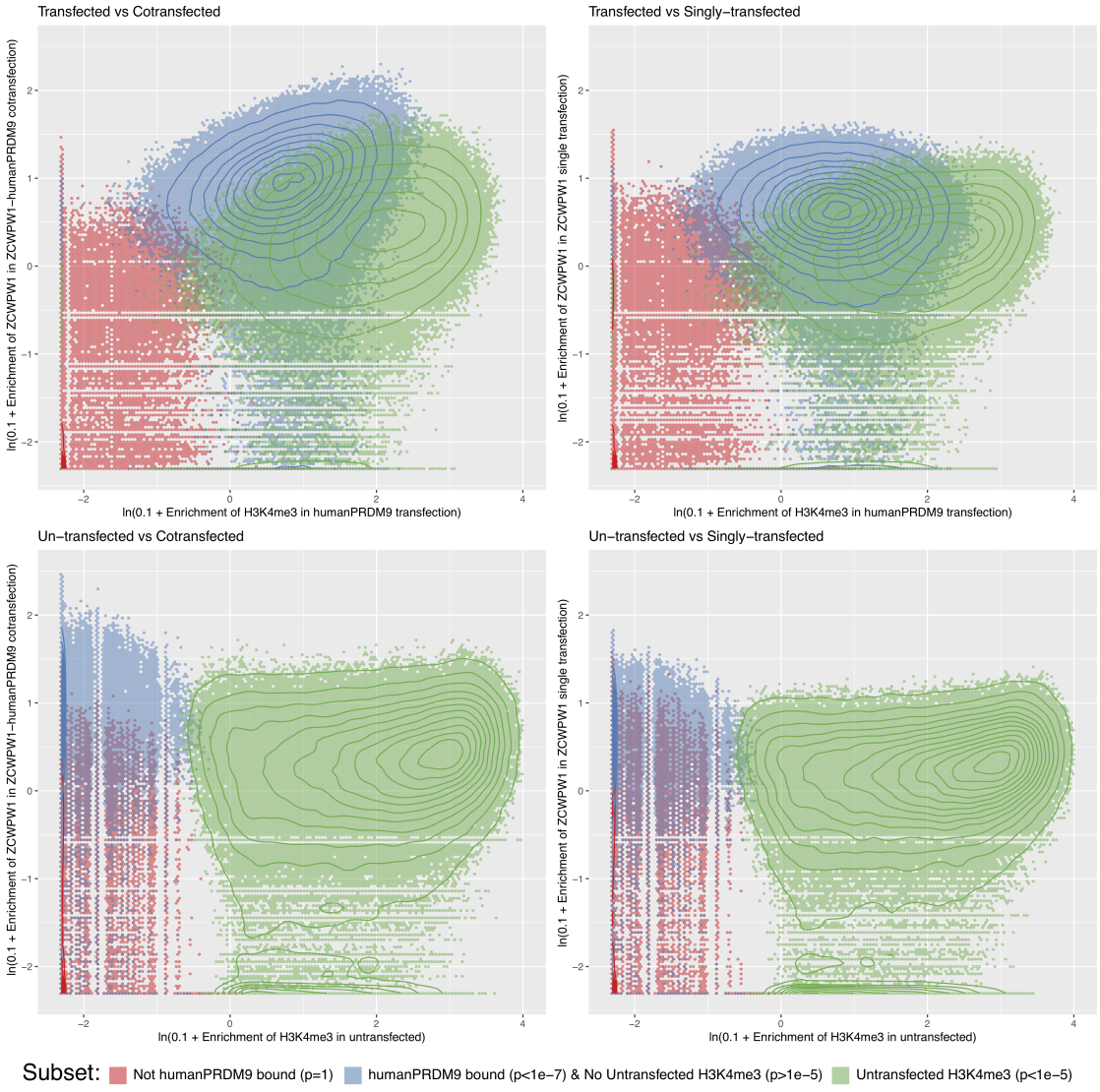


Figure A.11: Dependence of ZCWPW1 enrichment on H3K4me3. Subplot titles describe human PRDM9 transfection status vs ZCWPW1 cotransfection (with human PRDM9) status. Enrichment was force called in 100bp windows across the genome. Data for each colour group was downsampled to have the same number of data points per group. Hexagons are coloured if at least 5 data points fall within that bin. Solid lines show density contours estimated by MASS::kde2d() in R.

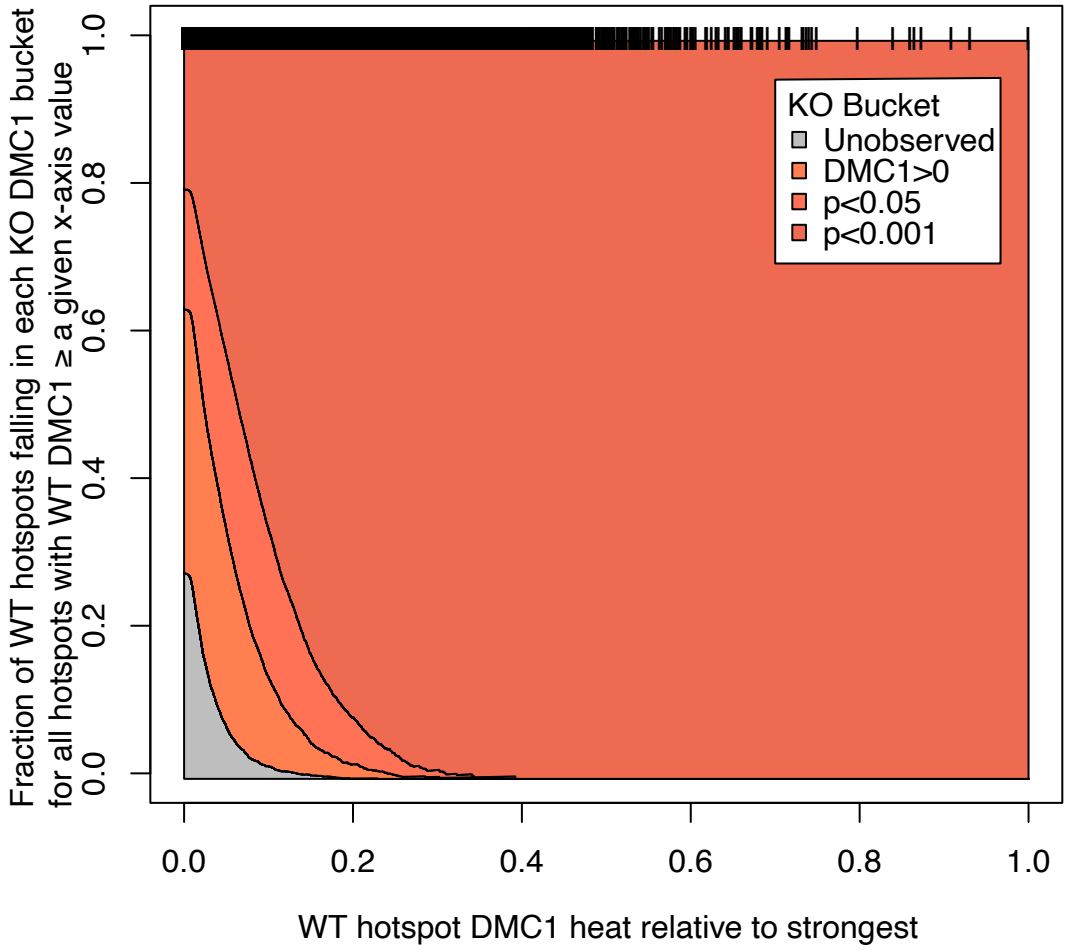


Figure A.12: Fraction of WT hotspot locations seen in *Zcwpw1*$^{-/-}$ DMC1 ChIP-seq at different p-values. Black bars along the top of the plot show the heat of individual hotspots relative to the hottest, according to the DMC1 data, in the WT male mouse. Y-axis values at x = 0 show the fraction of all hotspots falling into the buckets shown in the inset colour legend. As the x-axis increases the y-axis values show the same thing, but only for those hotspots with a heat greater than or equal to the x-axis value, that is those black bars further to the right. Therefore, almost all WT hotspots with activity >20% of the hottest hotspot are observed, and non-observed hotspots show only weak activity in WT, and so our power to detect them is expected to be reduced. 'DMC1>0' refers to the hotspot locations at which DMC1 signal is observed in *Zcwpw1*$^{-/-}$ DMC1 ChIP-seq, but with significance level (p-value) greater than or equal to 0.05, 'p<0.05' refers to the locations at which this significance level is less than 0.05 but greater or equal to 0.001, and 'p<0.001' refers to locations at which the p-value is less than 0.001.

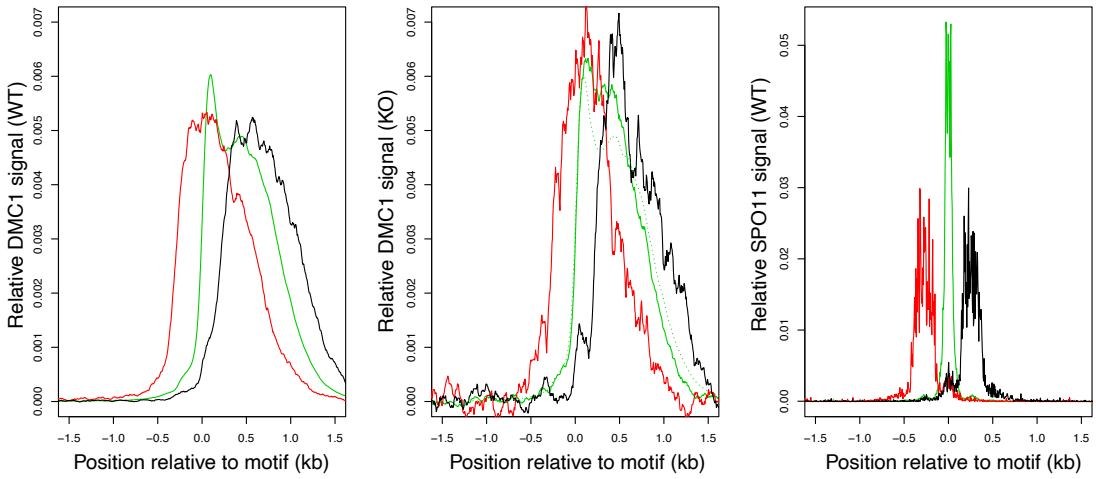


Figure A.13: DSBs in *Zcwpw1*$^{-/-}$ are positioned at WT locations within hotspots. DMC1 signal (left and center) is stratified by SPO11 (right). Stratification is into active hotspots (top 30%) with >90% of the SPO11 signal in the central 300bp (green), and <50% central, and >90% upstream of the PRDM9 binding motif (red) or <10% upstream (black).

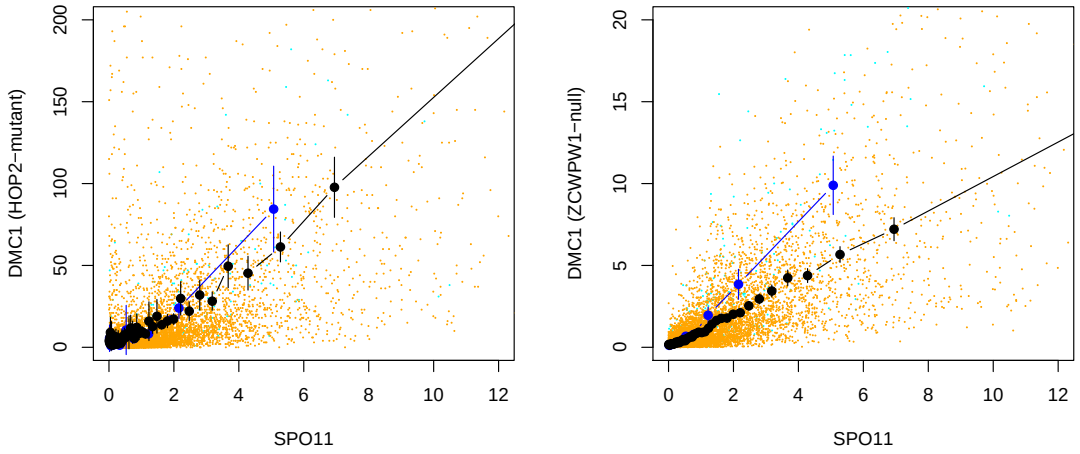


Figure A.14: *Hop2*$^{-/-}$ and *Zcwpw1*$^{-/-}$ mouse knockout mutants show the same linear relationship of DMC1 ChIP-seq vs SPO11 (WT). DMC1 data from *Hop2*$^{-/-}$ mice is from GSM851661 (Khil et al., 2012).

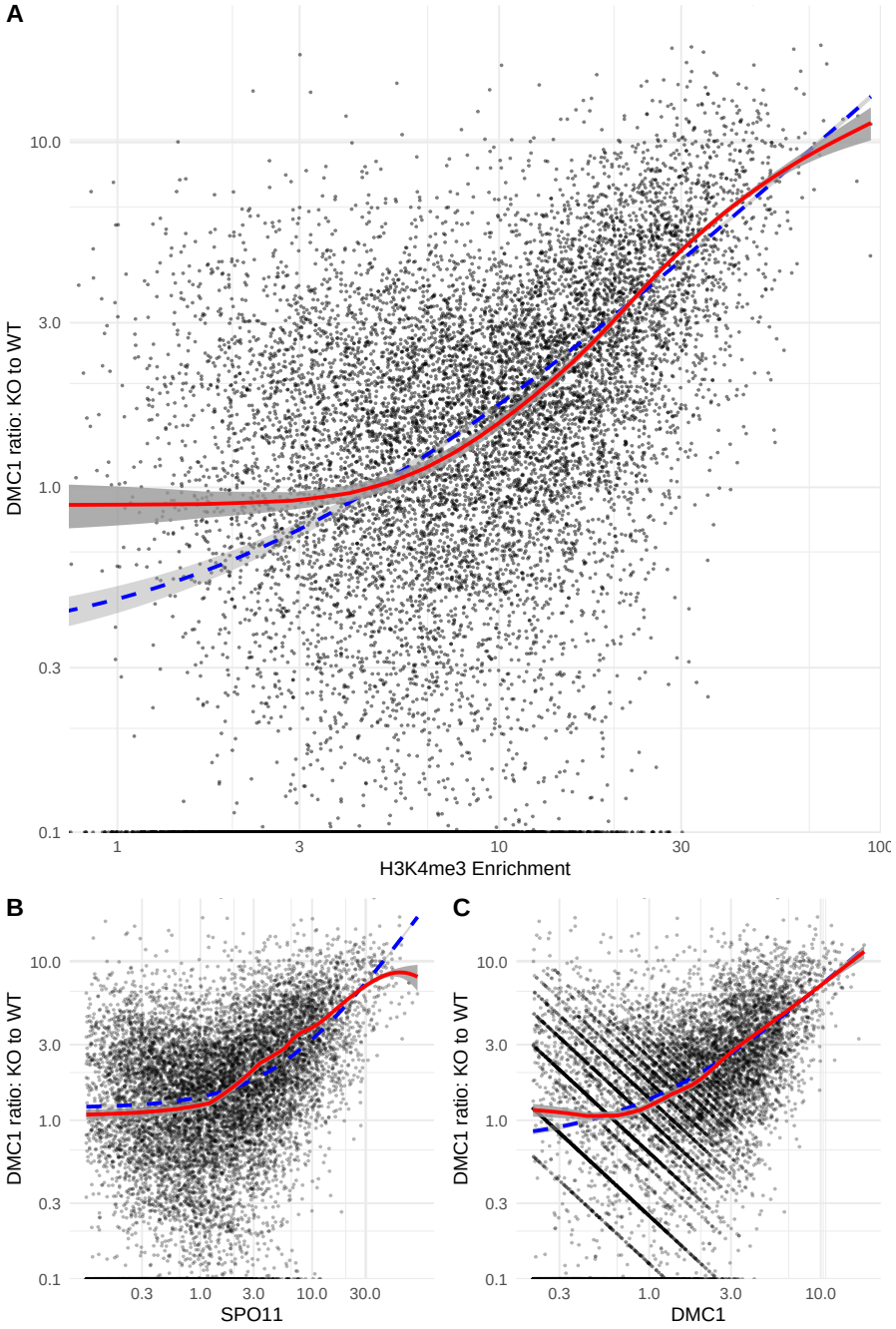


Figure A.15: Regression of DMC1 ratio on H3K4me3, SPO11, and DMC1. The DMC1 signal in the KO relative to the WT increases as H3K4me3 (PRDM9) increases. We calculated the ratio of KO to WT DMC1 force-called enrichment at each autosomal B6 mouse hotspot not overlapping pre-existing H3K4me3. We excluded weak hotspots whose estimated SPO11 heat (within 2.5kb) or DMC1 WT heats were in the bottom 33% (because accurate ratio estimation is not possible for these hotspots). Dots: the force-called signal strength (of either H3K4me3, SPO11 or WT DCM1), vs the ratio, for each of the resulting hotspots. Blue dashed line, linear regression line of best fit (fit in linear space, displayed in log space). Red line: Generalised Additive model (able to fit non-linear effects if present, again fit in linear space).

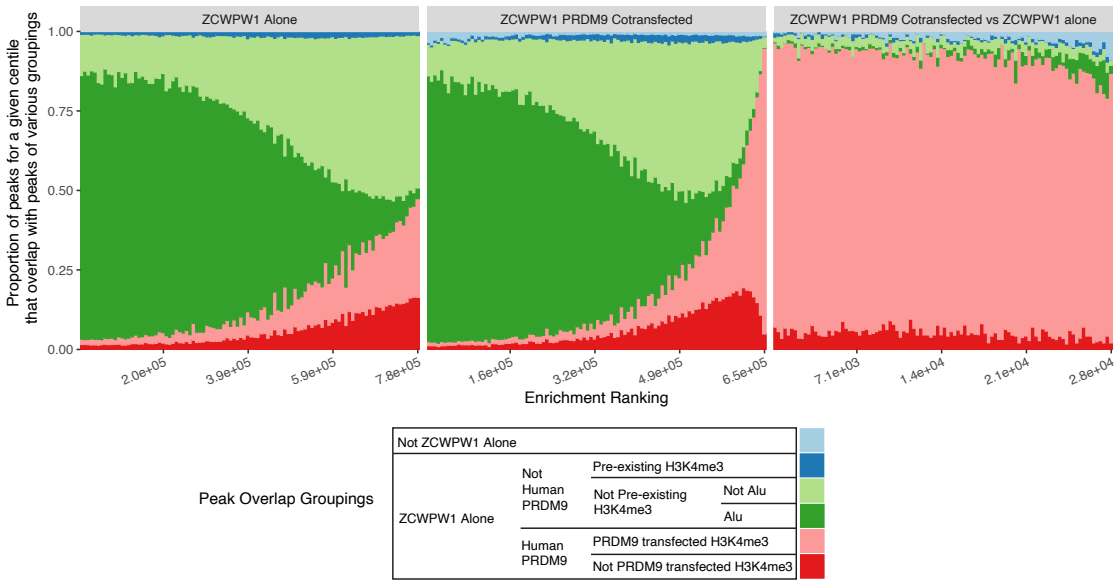


Figure A.16: Proportion of ZCWPW1 peaks overlapping various other marks, ordered by enrichment of ZCWPW1 binding over input. For example dark green peaks are those which overlap with ZCWPW1 peaks when transfected alone, but not overlapping Human PRDM9 peaks, and not overlapping pre-existing H3K4me3 peaks but do overlap with Alu repeats.

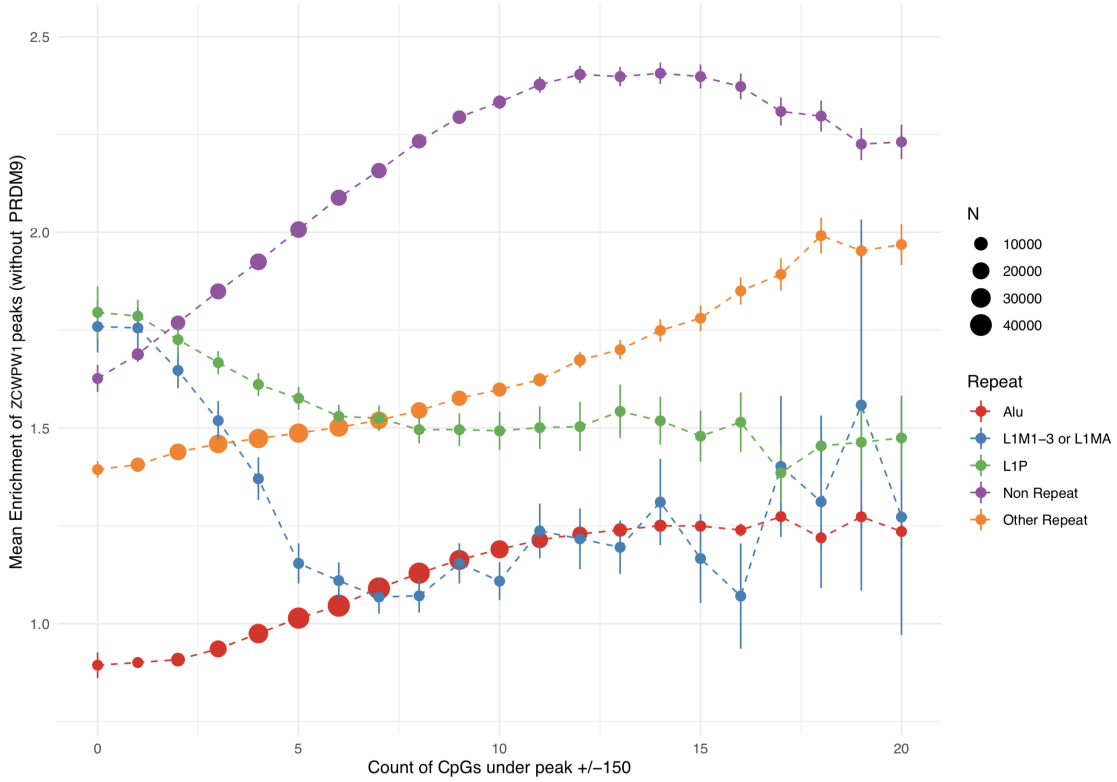


Figure A.17: CpG count around ZCWPW1 peaks (+/-150bp, with input coverage>5) is positively associated with ZCWPW1 enrichment in both peaks overlapping Alus and peaks not overlapping Alus, but not at L1M1-3, L1MA or L1P repeats. Error bars show ±2 s.e. of the mean.

What chromosomes are these which unite to form pairs? Does a paternal chromosome unite with a paternal and a maternal with a maternal, or does a paternal chromosome unite with a maternal one? The following considerations show that the latter view is probably the true one.

— Montgomery, 1901, p221

References

Weismann, A. (1887). "Ueber Die Zahl Der Richtungskörper Und Über Ihre Bedeutung Für Die Vererbung (On the Number of Polar Bodies and Their Significance in Heredity)". In: *Jena G. Fischer*, 34 (358 in translation) (cit. on p. 1).

Ohkura, H. (May 2015). "Meiosis: An Overview of Key Differences from Mitosis". In: *Cold Spring Harbor Perspectives in Biology* 7.5, a015859+. DOI: 10.1101/cshperspect.a015859. pmid: 25605710 (cit. on p. 2).

White, M. J. D. (1978). *Modes of Speciation*. San Francisco: W.H. Freeman (cit. on p. 2).

Debortoli, N. et al. (Mar. 21, 2016). "Genetic Exchange among Bdelloid Rotifers Is More Likely Due to Horizontal Gene Transfer Than to Meiotic Sex". In: *Current Biology* 26.6, pp. 723–732. DOI: 10.1016/j.cub.2016.01.031 (cit. on p. 2).

Laine, V. N., T. Sackton, and M. Meselson (Aug. 10, 2020). "Sexual Reproduction in Bdelloid Rotifers". In: *bioRxiv*, p. 2020.08.06.239590. DOI: 10.1101/2020.08.06.239590 (cit. on p. 2).

Bernstein, H. and C. Bernstein (Sept. 11, 2013). "Evolutionary Origin and Adaptive Function of Meiosis". In: *Meiosis*. DOI: 10.5772/56557 (cit. on p. 2).

Shao, Y. et al. (Aug. 2018). "Creating a Functional Single-Chromosome Yeast". In: *Nature* 560.7718, pp. 331–335. DOI: 10.1038/s41586-018-0382-x (cit. on p. 3).

Davies, B. et al. (Feb. 2016). "Re-Engineering the Zinc Fingers of PRDM9 Reverses Hybrid Sterility in Mice". In: *Nature* 530.7589, pp. 171–176. DOI: 10.1038/nature16931. pmid: 26840484 (cit. on pp. 3, 24, 84, 92, 112, 118, 124, 127–129, 132).

Hassold, T., H. Hall, and P. Hunt (July 2007). "The Origin of Human Aneuploidy: Where We Have Been, Where We Are Going". In: *Human Molecular Genetics* 16.R2, R203–R208. DOI: 10.1093/hmg/ddm243. pmid: 17911163 (cit. on p. 3).

Datta, J. et al. (Sept. 1, 2016). "Prevalence of Infertility and Help Seeking among 15 000 Women and Men". In: *Human Reproduction* 31.9, pp. 2108–2118. DOI: 10.1093/humrep/dew123 (cit. on p. 3).

Hamada, A., S. C. Esteves, and A. Agarwal (Apr. 2011). "Unexplained Male Infertility: Potential Causes and Management". In: *Human Andrology* 1.1, p. 2. DOI: 10.1097/01.XHA.0000397686.82729.09 (cit. on p. 3).

Magnus, M. C. et al. (Mar. 20, 2019). "Role of Maternal Age and Pregnancy History in Risk of Miscarriage: Prospective Register Based Study". In: *BMJ* 364, p. l869. DOI: 10.1136/bmj.l869. pmid: 30894356 (cit. on p. 3).

Myers, S. et al. (Sept. 2008). "A Common Sequence Motif Associated with Recombination Hot Spots and Genome Instability in Humans". In: *Nature Genetics* 40.9, pp. 1124–1129. DOI: 10.1038/ng.213 (cit. on pp. 3, 23).

Zhou, Q. et al. (Mar. 2016). "Complete Meiosis from Embryonic Stem Cell-Derived Germ Cells In Vitro". In: *Cell stem cell* 18.3, pp. 330–340. DOI: 10.1016/j.stem.2016.01.017. pmid: 26923202 (cit. on pp. 3, 29).

Cleland, J. et al. (July 14, 2012). "Contraception and Health". In: *The Lancet* 380.9837, pp. 149–156. DOI: 10.1016/S0140-6736(12)60609-6 (cit. on p. 3).

Schultz, N., F. K. Hamra, and D. L. Garbers (Oct. 2003). "A Multitude of Genes Expressed Solely in Meiotic or Postmeiotic Spermatogenic Cells Offers a Myriad of Contraceptive Targets". In: *Proceedings of the National Academy of Sciences* 100.21, pp. 12201–12206. DOI: 10.1073/pnas.1635054100. pmid: 14526100 (cit. on pp. 3, 78).

Djureinovic, D. et al. (June 2014). "The Human Testis-Specific Proteome Defined by Transcriptomics and Antibody-Based Profiling". In: *Molecular human reproduction* 20.6, pp. 476–488. DOI: 10.1093/molehr/gau018. pmid: 24598113 (cit. on pp. 3, 4, 29).

Mele, M. et al. (May 2015). "The Human Transcriptome across Tissues and Individuals". In: *Science* 348.6235, pp. 660–665. DOI: 10.1126/science.aaa0355. pmid: 25954002 (cit. on pp. 3, 28).

Uhlén, M. et al. (Jan. 2015). "Tissue-Based Map of the Human Proteome". In: *Science* 347.6220, p. 1260419. DOI: 10.1126/science.1260419. pmid: 25613900 (cit. on pp. 3, 4, 28, 96).

Uhlén, M. et al. (Apr. 2016). "Transcriptomics Resources of Human Tissues and Organs". In: *Molecular Systems Biology* 12.4, p. 862. DOI: 10.15252/msb.20155865. pmid: 27044256 (cit. on pp. 4, 28).

The Human Protein Atlas (2019). *The Human Tissue Specific Proteome* (cit. on p. 4).

Guo, J. H. et al. (Aug. 2005). "Transcriptomic Analyses Support the Similarity of Gene Expression between Brain and Testis in Human as Well as Mouse". In: *Cytogenetic and Genome Research* 111.2, pp. 107–109. DOI: 10.1159/000086378. pmid: 16103650 (cit. on p. 4).

Han, X. et al. (Feb. 22, 2018). "Mapping the Mouse Cell Atlas by Microwell-Seq". In: *Cell* 172.5, 1091–1107.e17. DOI: 10.1016/j.cell.2018.02.001. pmid: 29474909 (cit. on pp. 4, 7, 11, 134).

Fijak, M. and A. Meinhardt (2006). "The Testis in Immune Privilege". In: *Immunological Reviews* 213.1, pp. 66–81. DOI: 10.1111/j.1600-065X.2006.00438.x (cit. on pp. 4, 63).

Whitehurst, A. W. (2014). "Cause and Consequence of Cancer/Testis Antigen Activation in Cancer". In: *Annual Review of Pharmacology and Toxicology* 54.1, pp. 251–272. DOI: 10.1146/annurev-pharmtox-011112-140326. pmid: 24160706 (cit. on p. 4).

Rathke, C. et al. (Mar. 1, 2014). "Chromatin Dynamics during Spermiogenesis". In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*. Chromatin and Epigenetic Regulation of Animal Development 1839.3, pp. 155–168. DOI: 10.1016/j.bbagrm.2013.08.004 (cit. on p. 4).

Kleene, K. C. (July 1, 2013). "Connecting Cis-Elements and Trans-Factors with Mechanisms of Developmental Regulation of mRNA Translation in Meiotic and Haploid Mammalian Spermatogenic Cells". In: *Reproduction* 146.1, R1–R19. DOI: 10.1530/REP-12-0362 (cit. on p. 4).

Kaupp, U. B., N. D. Kashikar, and I. Weyand (2008). "Mechanisms of Sperm Chemotaxis". In: *Annual Review of Physiology* 70.1, pp. 93–117. DOI: 10.1146/annurev.physiol.70.113006.100654. pmid: 17988206 (cit. on p. 5).

Thompson, L. A. et al. (1992). "The Leukocytic Reaction of the Human Uterine Cervix". In: *American Journal of Reproductive Immunology* 28.2, pp. 85–89. DOI: 10.1111/j.1600-0897.1992.tb00765.x (cit. on p. 5).

Gould, J. E., J. W. Overstreet, and F. W. Hanson (Dec. 1, 1984). "Assessment of Human Sperm Function After Recovery from the Female Reproductive Tract". In: *Biology of Reproduction* 31.5, pp. 888–894. DOI: 10.1095/biolreprod31.5.888 (cit. on p. 5).

Wilcox, A. J., C. R. Weinberg, and D. D. Baird (Dec. 7, 1995). "Timing of Sexual Intercourse in Relation to Ovulation — Effects on the Probability of Conception, Survival of the Pregnancy, and Sex of the Baby". In: *New England Journal of Medicine* 333.23, pp. 1517–1521. DOI: 10.1056/NEJM199512073332301. pmid: 7477165 (cit. on p. 5).

Kong, A. et al. (Aug. 2012). "Rate of de Novo Mutations and the Importance of Father's Age to Disease Risk". In: *Nature* 488.7412, pp. 471–475. DOI: 10.1038/nature11396 (cit. on p. 5).

Michaelson, J. J. et al. (Dec. 21, 2012). "Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation". In: *Cell* 151.7, pp. 1431–1442. DOI: 10.1016/j.cell.2012.11.019. pmid: 23260136 (cit. on p. 5).

Goldmann, J. M. et al. (Aug. 2016). "Parent-of-Origin-Specific Signatures of de Novo Mutations". In: *Nature Genetics* 48.8, pp. 935–939. DOI: 10.1038/ng.3597 (cit. on p. 5).

Rahbari, R. et al. (Feb. 2016). "Timing, Rates and Spectra of Human Germline Mutation". In: *Nature Genetics* 48.2, pp. 126–133. DOI: 10.1038/ng.3469 (cit. on p. 5).

Jónsson, H. et al. (Sept. 2017). "Parental Influence on Human Germline de Novo Mutations in 1,548 Trios from Iceland". In: *Nature* 549.7673, pp. 519–522. DOI: 10.1038/nature24018 (cit. on p. 5).

Grimaldi, D. A. and M. S. Engel (Sept. 2007). "Why Descriptive Science Still Matters". In: *BioScience* 57.8, pp. 646–647. DOI: 10.1641/B570802 (cit. on p. 5).

Wood, V. et al. (Feb. 20, 2019). "Hidden in Plain Sight: What Remains to Be Discovered in the Eukaryotic Proteome?" In: *Open Biology* 9.2, p. 180241. DOI: 10.1098/rsob.180241 (cit. on p. 5).

Edwards, A. M. et al. (Feb. 2011). "Too Many Roads Not Taken". In: *Nature* 470.7333, pp. 163–165. DOI: 10.1038/470163a (cit. on p. 5).

Stoeger, T. et al. (Sept. 18, 2018). "Large-Scale Investigation of the Reasons Why Potentially Important Genes Are Ignored". In: *PLOS Biology* 16.9, e2006643. DOI: 10.1371/journal.pbio.2006643 (cit. on p. 6).

Haynes, W. A., A. Tomczak, and P. Khatri (Jan. 22, 2018). "Gene Annotation Bias Impedes Biomedical Research". In: *Scientific Reports* 8.1, pp. 1–7. DOI: 10.1038/s41598-018-19333-x (cit. on p. 6).

Genome List - Genome - NCBI (2019). URL: https://www.ncbi.nlm.nih.gov/genome/browse/#!/eukaryotes/ (visited on 12/03/2019) (cit. on p. 6).

Gajbhiye, R., J. N. Fung, and G. W. Montgomery (Oct. 12, 2018). "Complex Genetics of Female Fertility". In: *npj Genomic Medicine* 3.1, pp. 1–10. DOI: 10.1038/s41525-018-0068-1 (cit. on p. 6).

Visscher, P. M. et al. (July 6, 2017). "10 Years of GWAS Discovery: Biology, Function, and Translation". In: *The American Journal of Human Genetics* 101.1, pp. 5–22. DOI: 10.1016/j.ajhg.2017.06.005 (cit. on p. 6).

Gallagher, M. D. and A. S. Chen-Plotkin (May 3, 2018). "The Post-GWAS Era: From Association to Function". In: *American Journal of Human Genetics* 102.5, pp. 717–730. DOI: 10.1016/j.ajhg.2018.04.002. pmid: 29727686 (cit. on p. 6).

Struck, T. J., B. K. Mannakee, and R. N. Gutenkunst (Aug. 13, 2018). "The Impact of Genome-Wide Association Studies on Biomedical Research Publications". In: *Human Genomics* 12.1, p. 38. DOI: 10.1186/s40246-018-0172-4 (cit. on p. 6).

"The Knockout Mouse Project" (Sept. 2004). In: *Nature Genetics* 36.9, pp. 921–924. DOI: 10.1038/ng0904-921 (cit. on p. 6).

Meehan, T. F. et al. (Aug. 2017). "Disease Model Discovery from 3,328 Gene Knockouts by The International Mouse Phenotyping Consortium". In: *Nature Genetics* 49.8, pp. 1231–1238. DOI: 10.1038/ng.3901 (cit. on p. 6).

Birling, M.-C. et al. (Nov. 22, 2019). "A Resource of Targeted Mutant Mouse Lines for 5,061 Genes". In: *bioRxiv*, p. 844092. DOI: 10.1101/844092 (cit. on p. 6).

Illergård, K., D. H. Ardell, and A. Elofsson (2009). "Structure Is Three to Ten Times More Conserved than Sequence—A Study of Structural Response in Protein Cores". In: *Proteins: Structure, Function, and Bioinformatics* 77.3, pp. 499–508. DOI: 10.1002/prot.22458 (cit. on p. 7).

Sousounis, K. et al. (Aug. 2, 2012). "Conservation of the Three-Dimensional Structure in Non-Homologous or Unrelated Proteins". In: *Human Genomics* 6.1, p. 10. DOI: 10.1186/1479-7364-6-10. pmid: 23244440 (cit. on p. 7).

wwPDB consortium (Jan. 8, 2019). "Protein Data Bank: The Single Global Archive for 3D Macromolecular Structure Data". In: *Nucleic Acids Research* 47 (Database issue), pp. D520–D528. DOI: 10.1093/nar/gky949. pmid: 30357364 (cit. on p. 7).

Grabowski, M. et al. (Mar. 2016). "The Impact of Structural Genomics: The First Quindecennial". In: *Journal of structural and functional genomics* 17.1, pp. 1–16. DOI: 10.1007/s10969-016-9201-5. pmid: 26935210 (cit. on p. 7).

Khafizov, K. et al. (Mar. 11, 2014). "Trends in Structural Coverage of the Protein Universe and the Impact of the Protein Structure Initiative". In: *Proceedings of the National Academy of Sciences* 111.10, pp. 3733–3738. DOI: 10.1073/pnas.1321614111. pmid: 24567391 (cit. on p. 7).

Regev, A. et al. (Dec. 5, 2017). "The Human Cell Atlas". In: *eLife* 6. Ed. by T. R. Gingeras, e27041. DOI: 10.7554/eLife.27041 (cit. on p. 7).

Regev, A. et al. (Oct. 11, 2018). "The Human Cell Atlas White Paper". In: arXiv: 1810.05192 [q-bio] (cit. on p. 7).

The Tabula Muris Consortium (Oct. 2018). "Single-Cell Transcriptomics of 20 Mouse Organs Creates a Tabula Muris". In: *Nature* 562.7727, pp. 367–372. DOI: 10.1038/s41586-018-0590-4 (cit. on p. 7).

Warner, J. R. (Nov. 1, 1999). "The Economics of Ribosome Biosynthesis in Yeast". In: *Trends in Biochemical Sciences* 24.11, pp. 437–440. DOI: 10.1016/S0968-0004(99)01460-7 (cit. on p. 7).

Roozemond, R. C. (Nov. 1, 1976). "Ultramicrochemical Determination of Nucleic Acids in Individual Cells Using the Zeiss UMSP-I Microspectrophotometer. Application to Isolated Rat Hepatocytes of Different Ploidy Classes". In: *The Histochemical Journal* 8.6, pp. 625–638. DOI: 10.1007/BF01003963 (cit. on p. 7).

Uemura, E. (Mar. 1, 1980). "Age-Related Changes in Neuronal RNA Content in Rhesus Monkeys (Macaca Mulatta)". In: *Brain Research Bulletin* 5.2, pp. 117–119. DOI: 10.1016/0361-9230(80)90182-3 (cit. on p. 7).

Tang, F., K. Lao, and M. A. Surani (Apr. 2011). "Development and Applications of Single-Cell Transcriptome Analysis". In: *Nature Methods* 8.4, S6–S11. DOI: 10.1038/nmeth.1557 (cit. on p. 7).

Marguerat, S. et al. (Oct. 2012). "Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells". In: *Cell* 151.3, pp. 671–683. DOI: 10.1016/j.cell.2012.09.019. pmid: 23101633 (cit. on p. 7).

Macaulay, I. C. and T. Voet (Jan. 30, 2014). "Single Cell Genomics: Advances and Future Perspectives". In: *PLOS Genetics* 10.1, e1004126. DOI: 10.1371/journal.pgen.1004126 (cit. on p. 7).

Raj, A. et al. (Sept. 2006). "Stochastic mRNA Synthesis in Mammalian Cells". In: *PLOS Biology* 4.10, e309+. DOI: 10.1371/journal.pbio.0040309. pmid: 16848618 (cit. on p. 8).

Chubb, J. R. et al. (May 23, 2006). "Transcriptional Pulsing of a Developmental Gene". In: *Current Biology* 16.10, pp. 1018–1025. DOI: 10.1016/j.cub.2006.03.092 (cit. on p. 8).

Gall, J. G. and M. L. Pardue (June 1, 1969). "Formation and Detection of Rna-Dna Hybrid Molecules in Cytological Preparations". In: *Proceedings of the National Academy of Sciences* 63.2, pp. 378–383. DOI: 10.1073/pnas.63.2.378. pmid: 4895535 (cit. on p. 8).

Singer, R. H. and D. C. Ward (Dec. 1, 1982). "Actin Gene Expression Visualized in Chicken Muscle Tissue Culture by Using in Situ Hybridization with a Biotinated Nucleotide Analog". In: *Proceedings of the National Academy of Sciences* 79.23, pp. 7331–7335. DOI: 10.1073/pnas.79.23.7331. pmid: 6961411 (cit. on p. 8).

Kislauskis, E. H. et al. (Oct. 1, 1993). "Isoform-Specific 3'-Untranslated Sequences Sort Alpha-Cardiac and Beta-Cytoplasmic Actin Messenger RNAs to Different Cytoplasmic Compartments." In: *The Journal of Cell Biology* 123.1, pp. 165–172. DOI: 10.1083/jcb.123.1.165. pmid: 8408195 (cit. on p. 8).

Femino, A. M. et al. (Apr. 24, 1998). "Visualization of Single RNA Transcripts in Situ". In: *Science* 280.5363, pp. 585–590. DOI: 10.1126/science.280.5363.585. pmid: 9554849 (cit. on p. 8).

Raj, A. et al. (Sept. 2008). "Imaging Individual mRNA Molecules Using Multiple Singly Labeled Probes". In: *Nature Methods* 5.10, pp. 877–879. DOI: 10.1038/nmeth.1253. pmid: 18806792 (cit. on p. 8).

Lubeck, E. et al. (Apr. 2014). "Single-Cell in Situ RNA Profiling by Sequential Hybridization". In: *Nature Methods* 11.4, pp. 360–361. DOI: 10.1038/nmeth.2892 (cit. on p. 8).

Eng, C.-H. L. et al. (Dec. 2017). "Profiling the Transcriptome with RNA SPOTs". In: *Nature Methods* 14.12, pp. 1153–1155. DOI: 10.1038/nmeth.4500 (cit. on p. 8).

Eng, C.-H. L. et al. (Apr. 2019). "Transcriptome-Scale Super-Resolved Imaging in Tissues by RNA seqFISH+". In: *Nature* 568.7751, pp. 235–239. DOI: 10.1038/s41586-019-1049-y (cit. on p. 8).

Chen, K. H. et al. (Apr. 24, 2015). "Spatially Resolved, Highly Multiplexed RNA Profiling in Single Cells". In: *Science* 348.6233. DOI: 10.1126/science.aaa6090. pmid: 25858977 (cit. on p. 8).

Moffitt, J. R., J. Hao, G. Wang, et al. (Sept. 27, 2016). "High-Throughput Single-Cell Gene-Expression Profiling with Multiplexed Error-Robust Fluorescence in Situ Hybridization". In: *Proceedings of the National Academy of Sciences* 113.39, pp. 11046–11051. DOI: 10.1073/pnas.1612826113. pmid: 27625426 (cit. on p. 8).

Moffitt, J. R., J. Hao, D. Bambah-Mukku, et al. (Dec. 2016). "High-Performance Multiplexed Fluorescence in Situ Hybridization in Culture and Tissue with Matrix

Imprinting and Clearing". In: *Proceedings of the National Academy of Sciences* 113.50, pp. 14456–14461. DOI: 10.1073/pnas.1617699113. pmid: 27911841 (cit. on p. 8).

Xia, C., H. P. Babcock, et al. (May 22, 2019). "Multiplexed Detection of RNA Using MERFISH and Branched DNA Amplification". In: *Scientific Reports* 9.1, pp. 1–13. DOI: 10.1038/s41598-019-43943-8 (cit. on p. 8).

Shah, S. et al. (July 12, 2018). "Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH". In: *Cell* 174.2, 363–376.e16. DOI: 10.1016/j.cell.2018.05.035. pmid: 29887381 (cit. on p. 8).

Xia, C., J. Fan, et al. (Sept. 24, 2019). "Spatial Transcriptome Profiling by MERFISH Reveals Subcellular RNA Compartmentalization and Cell Cycle-Dependent Gene Expression". In: *Proceedings of the National Academy of Sciences* 116.39, pp. 19490–19499. DOI: 10.1073/pnas.1912459116. pmid: 31501331 (cit. on p. 8).

Gibson, U. E., C. A. Heid, and P. M. Williams (Jan. 10, 1996). "A Novel Method for Real Time Quantitative RT-PCR." In: *Genome Research* 6.10, pp. 995–1001. DOI: 10.1101/gr.6.10.995. pmid: 8908519 (cit. on p. 8).

Heid, C. A. et al. (Jan. 10, 1996). "Real Time Quantitative PCR." In: *Genome Research* 6.10, pp. 986–994. DOI: 10.1101/gr.6.10.986. pmid: 8908518 (cit. on p. 8).

Chiang, P. W. et al. (Jan. 10, 1996). "Use of a Fluorescent-PCR Reaction to Detect Genomic Sequence Copy Number and Transcriptional Abundance." In: *Genome Research* 6.10, pp. 1013–1026. DOI: 10.1101/gr.6.10.1013. pmid: 8908521 (cit. on p. 8).

Palmer, S. et al. (Oct. 1, 2003). "New Real-Time Reverse Transcriptase-Initiated PCR Assay with Single-Copy Sensitivity for Human Immunodeficiency Virus Type 1 RNA in Plasma". In: *Journal of Clinical Microbiology* 41.10, pp. 4531–4536. DOI: 10.1128/JCM.41.10.4531-4536.2003. pmid: 14532178 (cit. on p. 9).

Wong, M. L. and J. F. Medrano (July 1, 2005). "Real-Time PCR for mRNA Quantitation". In: *BioTechniques* 39.1, pp. 75–85. DOI: 10.2144/05391RV01 (cit. on p. 9).

Bustin, S. A. et al. (Apr. 1, 2009). "The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments". In: *Clinical Chemistry* 55.4, pp. 611–622. DOI: 10.1373/clinchem.2008.112797. pmid: 19246619 (cit. on p. 9).

Alwine, J. C., D. J. Kemp, and G. R. Stark (Dec. 1, 1977). "Method for Detection of Specific RNAs in Agarose Gels by Transfer to Diazobenzyloxymethyl-Paper and Hybridization with DNA Probes". In: *Proceedings of the National Academy of Sciences* 74.12, pp. 5350–5354. DOI: 10.1073/pnas.74.12.5350. pmid: 414220 (cit. on p. 9).

Schena, M. et al. (Oct. 20, 1995). "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray". In: *Science* 270.5235, pp. 467–470. DOI: 10.1126/science.270.5235.467. pmid: 7569999 (cit. on p. 9).

DeRisi, J. et al. (Dec. 1996). "Use of a cDNA Microarray to Analyse Gene Expression Patterns in Human Cancer". In: *Nature Genetics* 14.4, pp. 457–460. DOI: 10.1038/ng1296-457. pmid: 8944026 (cit. on p. 9).

Schulze, A. and J. Downward (Aug. 2001). "Navigating Gene Expression Using Microarrays–a Technology Review". In: *Nature Cell Biology* 3.8, E190–195. DOI: 10.1038/35087138. pmid: 11483980 (cit. on p. 9).

Rustici, G. et al. (Sept. 2008). "Data Storage and Analysis in ArrayExpress and Expression Profiler". In: *Current Protocols in Bioinformatics* 23.1. DOI: 10.1002/0471250953.bi0713s23 (cit. on p. 9).

Mortazavi, A. et al. (July 2008). "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq". In: *Nature Methods* 5.7, pp. 621–628. DOI: 10.1038/nmeth.1226 (cit. on p. 9).

Cloonan, N. et al. (July 2008). "Stem Cell Transcriptome Profiling via Massive-Scale mRNA Sequencing". In: *Nature Methods* 5.7, pp. 613–619. DOI: 10.1038/nmeth.1223 (cit. on p. 9).

Wilhelm, B. T. et al. (June 2008). "Dynamic Repertoire of a Eukaryotic Transcriptome Surveyed at Single-Nucleotide Resolution". In: *Nature* 453.7199, pp. 1239–1243. DOI: 10.1038/nature07002 (cit. on p. 9).

Lister, R. et al. (May 2, 2008). "Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis". In: *Cell* 133.3, pp. 523–536. DOI: 10.1016/j.cell.2008.03.029 (cit. on p. 9).

Nagalakshmi, U. et al. (June 6, 2008). "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing". In: *Science* 320.5881, pp. 1344–1349. DOI: 10.1126/science.1158441. pmid: 18451266 (cit. on p. 9).

Zhao, S. et al. (Jan. 16, 2014). "Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells". In: *PLOS ONE* 9.1, e78644. DOI: 10.1371/journal.pone.0078644 (cit. on p. 10).

Stark, R., M. Grzelak, and J. Hadfield (Nov. 2019). "RNA Sequencing: The Teenage Years". In: *Nature Reviews Genetics* 20.11, pp. 631–656. DOI: 10.1038/s41576-019-0150-2 (cit. on p. 10).

Sharon, D. et al. (Nov. 2013). "A Single-Molecule Long-Read Survey of the Human Transcriptome". In: *Nature Biotechnology* 31.11, pp. 1009–1014. DOI: 10.1038/nbt.2705 (cit. on p. 10).

Cartolano, M. et al. (June 21, 2016). "cDNA Library Enrichment of Full Length Transcripts for SMRT Long Read Sequencing". In: *PLOS ONE* 11.6, e0157779. DOI: 10.1371/journal.pone.0157779 (cit. on p. 10).

Oikonomopoulos, S. et al. (Aug. 24, 2016). "Benchmarking of the Oxford Nanopore MinION Sequencing for Quantitative and Qualitative Assessment of cDNA Populations". In: *Scientific Reports* 6.1, pp. 1–13. DOI: 10.1038/srep31602 (cit. on p. 10).

Garalde, D. R. et al. (Mar. 2018). "Highly Parallel Direct RNA Sequencing on an Array of Nanopores". In: *Nature Methods* 15.3, pp. 201–206. DOI: 10.1038/nmeth.4577 (cit. on p. 10).

Workman, R. E. et al. (Dec. 2019). "Nanopore Native RNA Sequencing of a Human Poly(A) Transcriptome". In: *Nature Methods* 16.12, pp. 1297–1305. DOI: 10.1038/s41592-019-0617-2 (cit. on p. 10).

Tang, F. et al. (Apr. 2009). "mRNA-Seq Whole-Transcriptome Analysis of a Single Cell". In: *Nature Methods* 6.5, pp. 377–382. DOI: 10.1038/nmeth.1315. pmid: 19349980 (cit. on p. 10).

Islam, S. et al. (Apr. 2012). "Highly Multiplexed and Strand-Specific Single-Cell RNA 5[Prime] End Sequencing". In: *Nature Protocols* 7.5, pp. 813–828. DOI: 10.1038/nprot.2012.022 (cit. on p. 10).

Jaitin, D. A. et al. (Feb. 14, 2014). "Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types". In: *Science* 343.6172, pp. 776–779. DOI: 10.1126/science.1247651. pmid: 24531970 (cit. on p. 11).

Bose, S. et al. (June 6, 2015). "Scalable Microfluidics for Single-Cell RNA Printing and Sequencing". In: *Genome Biology* 16.1, p. 120. DOI: 10.1186/s13059-015-0684-3 (cit. on p. 11).

Fan, H. C., G. K. Fu, and S. P. A. Fodor (Feb. 6, 2015). "Combinatorial Labeling of Single Cells for Gene Expression Cytometry". In: *Science* 347.6222. DOI: 10.1126/science.1258367. pmid: 25657253 (cit. on p. 11).

Gierahn, T. M. et al. (Apr. 2017). "Seq-Well: Portable, Low-Cost RNA Sequencing of Single Cells at High Throughput". In: *Nature Methods* 14.4, pp. 395–398. DOI: 10.1038/nmeth.4179 (cit. on p. 11).

Macosko, E. Z. et al. (May 2015). "Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets". In: *Cell* 161.5, pp. 1202–1214. DOI: 10.1016/j.cell.2015.05.002 (cit. on pp. 11, 12, 39, 90).

Klein, A. M. et al. (May 21, 2015). "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells". In: *Cell* 161.5, pp. 1187–1201. DOI: 10.1016/j.cell.2015.04.044 (cit. on p. 11).

Saunders, A. et al. (Aug. 2018). "Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain". In: *Cell* 174.4, 1015–1030.e16. DOI: 10.1016/j.cell.2018.07.028. pmid: 30096299 (cit. on pp. 11, 31).

Cao, J. et al. (Feb. 2019). "The Single-Cell Transcriptional Landscape of Mammalian Organogenesis". In: *Nature* 566.7745, pp. 496–502. DOI: 10.1038/s41586-019-0969-x (cit. on p. 11).

Zheng, G. X. Y. et al. (Jan. 2017). "Massively Parallel Digital Transcriptional Profiling of Single Cells". In: *Nature Communications* 8, pp. 14049+. DOI: 10.1038/ncomms14049 (cit. on p. 11).

Svensson, V., R. Vento-Tormo, and S. A. Teichmann (Apr. 2018). "Exponential Scaling of Single-Cell RNA-Seq in the Past Decade". In: *Nature Protocols* 13.4, pp. 599–604. DOI: 10.1038/nprot.2017.149 (cit. on p. 11).

Svensson, V., E. d. V. Beltrame, and L. Pachter (Oct. 17, 2019). "A Curated Database Reveals Trends in Single-Cell Transcriptomics". In: *bioRxiv*, p. 742304. DOI: 10.1101/742304 (cit. on pp. 11, 30).

Best, K. et al. (Oct. 13, 2015). "Computational Analysis of Stochastic Heterogeneity in PCR Amplification Efficiency Revealed by Single Molecule Barcoding". In: *Scientific Reports* 5.1, pp. 1–13. DOI: 10.1038/srep14629 (cit. on p. 11).

Kivioja, T. et al. (Jan. 2012). "Counting Absolute Numbers of Molecules Using Unique Molecular Identifiers". In: *Nature Methods* 9.1, pp. 72–74. DOI: 10.1038/nmeth.1778 (cit. on p. 12).

Islam, S. et al. (Feb. 2014). "Quantitative Single-Cell RNA-Seq with Unique Molecular Identifiers". In: *Nature Methods* 11.2, pp. 163–166. DOI: 10.1038/nmeth.2772 (cit. on p. 12).

Kou, R. et al. (Jan. 11, 2016). "Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations". In: *PLOS ONE* 11.1, e0146638. DOI: 10.1371/journal.pone.0146638 (cit. on p. 12).

Wan, L. et al. (Sept. 2012). "Modeling RNA Degradation for RNA-Seq with Applications". In: *Biostatistics (Oxford, England)* 13.4, pp. 734–747. DOI: 10.1093/biostatistics/kxs001. pmid: 22353193 (cit. on p. 12).

Lahens, N. F. et al. (June 30, 2014). "IVT-Seq Reveals Extreme Bias in RNA Sequencing". In: *Genome Biology* 15.6, R86. DOI: 10.1186/gb-2014-15-6-r86 (cit. on p. 12).

Nam, D. K. et al. (Apr. 30, 2002). "Oligo(dT) Primer Generates a High Frequency of Truncated cDNAs through Internal Poly(A) Priming during Reverse Transcription". In: *Proceedings of the National Academy of Sciences* 99.9, pp. 6152–6156. DOI: 10.1073/pnas.092140899. pmid: 11972056 (cit. on p. 12).

Ntranos, V. et al. (Feb. 2019). "A Discriminative Learning Approach to Differential Expression Analysis for Single-Cell RNA-Seq". In: *Nature Methods* 16.2, pp. 163–166. DOI: 10.1038/s41592-018-0303-9 (cit. on p. 12).

Kang, H. M. et al. (Jan. 2018). "Multiplexed Droplet Single-Cell RNA-Sequencing Using Natural Genetic Variation". In: *Nature Biotechnology* 36.1, pp. 89–94. DOI: 10.1038/nbt.4042 (cit. on p. 12).

Stoeckius, M. et al. (Dec. 19, 2018). "Cell Hashing with Barcoded Antibodies Enables Multiplexing and Doublet Detection for Single Cell Genomics". In: *Genome Biology* 19.1, p. 224. DOI: 10.1186/s13059-018-1603-1 (cit. on p. 12).

Datlinger, P. et al. (Dec. 18, 2019). "Ultra-High Throughput Single-Cell RNA Sequencing by Combinatorial Fluidic Indexing". In: *bioRxiv*, p. 2019.12.17.879304. DOI: 10.1101/2019.12.17.879304 (cit. on p. 12).

Wolock, S. L., R. Lopez, and A. M. Klein (Apr. 24, 2019). "Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data". In: *Cell Systems* 8.4, 281–291.e9. DOI: 10.1016/j.cels.2018.11.005. pmid: 30954476 (cit. on p. 13).

McGinnis, C. S., L. M. Murrow, and Z. J. Gartner (Apr. 24, 2019). "DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors". In: *Cell Systems* 8.4, 329–337.e4. DOI: 10.1016/j.cels.2019.03.003. pmid: 30954475 (cit. on p. 13).

Lee, J. H. et al. (Mar. 21, 2014). "Highly Multiplexed Subcellular RNA Sequencing in Situ". In: *Science* 343.6177, pp. 1360–1363. DOI: 10.1126/science.1250212. pmid: 24578530 (cit. on p. 13).

Wang, X. et al. (July 27, 2018). "Three-Dimensional Intact-Tissue Sequencing of Single-Cell Transcriptional States". In: *Science* 361.6400. DOI: 10.1126/science.aat5691. pmid: 29930089 (cit. on p. 13).

Qian, X. et al. (Oct. 1, 2018). "A Spatial Atlas of Inhibitory Cell Types in Mouse Hippocampus". In: *bioRxiv*, p. 431957. DOI: 10.1101/431957 (cit. on p. 13).

Ståhl, P. L. et al. (July 1, 2016). "Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics". In: *Science* 353.6294, pp. 78–82. DOI: 10.1126/science.aaf2403. pmid: 27365449 (cit. on p. 13).

Rodriques, S. G. et al. (Mar. 29, 2019). "Slide-Seq: A Scalable Technology for Measuring Genome-Wide Expression at High Spatial Resolution". In: *Science* 363.6434, pp. 1463–1467. DOI: 10.1126/science.aaw1219. pmid: 30923225 (cit. on p. 13).

Vickovic, S. et al. (Oct. 2019). "High-Definition Spatial Transcriptomics for in Situ Tissue Profiling". In: *Nature Methods* 16.10, pp. 987–990. DOI: 10.1038/s41592-019-0548-y (cit. on p. 13).

Core, L. J., J. J. Waterfall, and J. T. Lis (Dec. 19, 2008). "Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters". In: *Science* 322.5909, pp. 1845–1848. DOI: 10.1126/science.1162228. pmid: 19056941 (cit. on p. 13).

Churchman, L. S. and J. S. Weissman (Jan. 2011). "Nascent Transcript Sequencing Visualizes Transcription at Nucleotide Resolution". In: *Nature* 469.7330, pp. 368–373. DOI: 10.1038/nature09652 (cit. on p. 13).

Mahat, D. B. et al. (Aug. 2016). "Base-Pair-Resolution Genome-Wide Mapping of Active RNA Polymerases Using Precision Nuclear Run-on (PRO-Seq)". In: *Nature Protocols* 11.8, pp. 1455–1476. DOI: 10.1038/nprot.2016.086 (cit. on p. 13).

Specht, H. and N. Slavov (Aug. 3, 2018). "Transformative Opportunities for Single-Cell Proteomics". In: *Journal of Proteome Research* 17.8, pp. 2565–2571. DOI: 10.1021/acs.jproteome.8b00257 (cit. on p. 13).

Marx, V. (Sept. 2019). "A Dream of Single-Cell Proteomics". In: *Nature Methods* 16.9, pp. 809–812. DOI: 10.1038/s41592-019-0540-6 (cit. on p. 14).

Liu, Y., A. Beyer, and R. Aebersold (Apr. 2016). "On the Dependency of Cellular Protein Levels on mRNA Abundance". In: *Cell* 165.3, pp. 535–550. DOI: 10.1016/j.cell.2016.03.014 (cit. on p. 14).

Oakberg, E. F. (Nov. 1956a). "Duration of Spermatogenesis in the Mouse and Timing of Stages of the Cycle of the Seminiferous Epithelium". In: *The American journal of anatomy* 99.3, pp. 507–516. DOI: 10.1002/aja.1000990307. pmid: 13402729 (cit. on pp. 15, 16).

Hess, R. A. and A. W. Vogl (Jan. 1, 2015). "1 - Sertoli Cell Anatomy and Cytoskeleton". In: *Sertoli Cell Biology (Second Edition)*. Ed. by M. D. Griswold. Oxford: Academic Press, pp. 1–55. DOI: 10.1016/B978-0-12-417047-6.00001-6 (cit. on p. 15).

Link, J. and V. Jantsch (Sept. 1, 2019). "Meiotic Chromosomes in Motion: A Perspective from Mus Musculus and Caenorhabditis Elegans". In: *Chromosoma* 128.3, pp. 317–330. DOI: 10.1007/s00412-019-00698-5 (cit. on p. 15).

Monesi, V. et al. (Jan. 1, 1978). "Chapter 2 Biochemistry of Male Germ Cell Differentiation in Mammals: RNA Synthesis in Meiotic and Postmeiotic Cells". In: *Current Topics in Developmental Biology*. Ed. by A. A. Moscona and A. Monroy. Vol. 12. Academic Press, pp. 11–36. DOI: 10.1016/S0070-2153(08)60592-X (cit. on pp. 15, 67).

Ahmed, E. A. and D. G. de Rooij (2009). "Staging of Mouse Seminiferous Tubule Cross-Sections". In: *Meiosis: Volume 2, Cytological Methods*. Ed. by S. Keeney. Methods in Molecular Biology. Totowa, NJ: Humana Press, pp. 263–277. DOI: 10.1007/978-1-60761-103-5_16 (cit. on p. 15).

Meistrich, M. L. and R. A. Hess (2013). "Assessment of Spermatogenesis Through Staging of Seminiferous Tubules". In: *Spermatogenesis: Methods and Protocols*. Ed. by D. T. Carrell and K. I. Aston. Methods in Molecular Biology. Totowa, NJ: Humana Press, pp. 299–307. DOI: 10.1007/978-1-62703-038-0_27 (cit. on p. 15).

Oakberg, E. F. (Nov. 1956b). "A Description of Spermiogenesis in the Mouse and Its Use in Analysis of the Cycle of the Seminiferous Epithelium and Germ Cell Renewal". In: *American Journal of Anatomy* 99.3, pp. 391–413. DOI: 10.1002/aja.1000990303 (cit. on pp. 15, 16).

Hess, R. A. and L. R. de Franca (2008). "Spermatogenesis and Cycle of the Seminiferous Epithelium". In: *Molecular Mechanisms in Spermatogenesis*. Ed. by C. Y. Cheng. Advances in Experimental Medicine and Biology. New York, NY: Springer New York, pp. 1–15. DOI: 10.1007/978-0-387-09597-4_1 (cit. on p. 15).

Nakata, H. et al. (Feb. 1, 2015). "Quantitative Analysis of the Cellular Composition in Seminiferous Tubules in Normal and Genetically Modified Infertile Mice". In: *Journal*

of Histochemistry & Cytochemistry 63.2, pp. 99–113. DOI: 10.1369/0022155414562045 (cit. on p. 15).

Muciaccia, B. et al. (Sept. 1, 2013). "Novel Stage Classification of Human Spermatogenesis Based on Acrosome Development". In: *Biology of Reproduction* 89.3. DOI: 10.1095/biolreprod.113.111682 (cit. on p. 15).

Osuru, H. P. et al. (2014). "The Acrosomal Protein SP-10 (Acrv1) Is an Ideal Marker for Staging of the Cycle of Seminiferous Epithelium in the Mouse". In: *Molecular Reproduction and Development* 81.10, pp. 896–907. DOI: 10.1002/mrd.22358 (cit. on pp. 15, 77).

Lara, N. L. M. et al. (Jan. 1, 2018). "Testis Physiology—Overview and Histology". In: *Encyclopedia of Reproduction (Second Edition)*. Ed. by M. K. Skinner. Oxford: Academic Press, pp. 105–116. DOI: 10.1016/B978-0-12-801238-3.64567-1 (cit. on p. 15).

Junqueira, L. and J. Carneiro (July 2005). *Basic Histology: Text & Atlas*. 11th ed. McGraw-Hill Medical. Fig 21.5+ (cit. on p. 15).

Clermont, Y. and M. Trott (Sept. 1, 1969). "Duration of the Cycle of the Seminiferous Epithelium in the Mouse and Hamster Determined by Means of 3H-Thymidine and Radioautography". In: *Fertility and Sterility* 20.5, pp. 805–817. DOI: 10.1016/S0015-0282(16)37153-9 (cit. on p. 16).

Heller, C. G., G. V. Heller, and M. J. Rowley (1969). "Human Spermatogenesis: An Estimate of the Duration of Each Cell Association and of Each Cell Type." In: (cit. on p. 16).

Leblond, C. P. and Y. Clermont (1952a). "Spermiogenesis of Rat, Mouse, Hamster and Guinea Pig as Revealed by the "Periodic Acid-Fuchsin Sulfurous Acid" Technique". In: *American Journal of Anatomy* 90.2, pp. 167–215. DOI: 10.1002/aja.1000900202 (cit. on p. 16).

— (1952b). "Definition of the Stages of the Cycle of the Seminiferous Epithelium in the Rat". In: *Annals of the New York Academy of Sciences* 55.4, pp. 548–573. DOI: 10.1111/j.1749-6632.1952.tb26576.x (cit. on p. 16).

Oakberg, E. F. (1971). "Spermatogonial Stem-Cell Renewal in the Mouse". In: *The Anatomical Record* 169.3, pp. 515–531. DOI: 10.1002/ar.1091690305 (cit. on p. 16).

Huckins, C. (1971). "The Spermatogonial Stem Cell Population in Adult Rats. I. Their Morphology, Proliferation and Maturation". In: *The Anatomical Record* 169.3, pp. 533–557. DOI: 10.1002/ar.1091690306 (cit. on p. 16).

Rooij, D. de (Mar. 1, 2001). "Proliferation and Differentiation of Spermatogonial Stem Cells". In: *Reproduction* 121.3, pp. 347–354. DOI: 10.1530/rep.0.1210347 (cit. on p. 16).

Rooij, D. de and L. D. Russell (2000). "All You Wanted to Know About Spermatogonia but Were Afraid to Ask". In: *Journal of Andrology* 21.6, pp. 776–798. DOI: 10.1002/j.1939-4640.2000.tb03408.x (cit. on p. 16).

Griswold, M. D. et al. (Feb. 2012). "Initiating Meiosis: The Case for Retinoic Acid". In: *Biology of Reproduction* 86.2. DOI: 10.1095/biolreprod.111.096610. pmid: 22075477 (cit. on p. 16).

Lin, Y. et al. (Dec. 2008). "Germ Cell-Intrinsic and -Extrinsic Factors Govern Meiotic Initiation in Mouse Embryos". In: *Science* 322.5908, pp. 1685–1687. DOI: 10.1126/science.1166340. pmid: 19074348 (cit. on p. 16).

Zhou, Q. et al. (Mar. 2008). "Expression of Stimulated by Retinoic Acid Gene 8 (Stra8) and Maturation of Murine Gonocytes and Spermatogonia Induced by Retinoic Acid

In Vitro1". In: *Biology of Reproduction* 78.3, pp. 537–545. DOI: 10.1095/biolreprod.107.064337 (cit. on pp. 16, 65).

Hogarth, C. A. et al. (Feb. 2015). "Processive Pulses of Retinoic Acid Propel Asynchronous and Continuous Murine Sperm Production". In: *Biology of Reproduction* 92.2. DOI: 10.1095/biolreprod.114.126326. pmid: 25519186 (cit. on p. 16).

Griswold, M. D. (Nov. 2015). "Spermatogenesis: The Commitment to Meiosis". In: *Physiological Reviews* 96.1, pp. 1–17. DOI: 10.1152/physrev.00013.2015 (cit. on p. 16).

Thompson, J. N. et al. (Feb. 18, 1964). "Vitamin A and Reproduction in Rats". In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 159.976, pp. 510–535. DOI: 10.1098/rspb.1964.0017 (cit. on p. 16).

De Winiwarter, H. von (1900). *Recherches sur l'Ovogenèse et l'Organogenèse de l'ovaire des Mammifères (Lapin et Homme)*. Liège: Vaillant-Carmanne (cit. on p. 17).

Grégoire, V. (1907). "La Formation Des Gemini Hétérotypiques Dans Les Végétaux." In: *Cellule* 24.367-420, p371 (cit. on p. 17).

Wilson, E. B. (1912). "Studies on Chromosomes. VIII. Observations on the Maturation-Phenomena in Certain Hemiptera and Other Forms, with Considerations on Synapsis and Reduction". In: *Journal of Experimental Zoology* 13.3, pp. 345–449. DOI: 10.1002/jez.1400130302 (cit. on p. 17).

Zickler, D. and N. Kleckner (Dec. 1, 1998). "The Leptotene-Zygotene Transition of Meiosis". In: *Annual Review of Genetics* 32.1, pp. 619–697. DOI: 10.1146/annurev.genet.32.1.619 (cit. on p. 17).

Moses, M. J. (Mar. 25, 1956). "Chromosomal Structures in Crayfish Spermatocytes". In: *The Journal of Biophysical and Biochemical Cytology* 2.2, pp. 215–218. DOI: 10.1083/jcb.2.2.215 (cit. on p. 17).

Fawcett, D. W. (July 25, 1956). "THE FINE STRUCTURE OF CHROMOSOMES IN THE MEIOTIC PROPHASE OF VERTEBRATE SPERMATOCYTES". In: *The Journal of Biophysical and Biochemical Cytology* 2.4, pp. 403–406. DOI: 10.1083/jcb.2.4.403 (cit. on p. 17).

Zickler, D. and N. Kleckner (May 2015). "Recombination, Pairing, and Synapsis of Homologs during Meiosis". In: *Cold Spring Harbor perspectives in biology* 7.6. DOI: 10.1101/cshperspect.a016626. pmid: 25986558 (cit. on pp. 17, 21).

Moens, P. B. (Dec. 1, 1968). "The Structure and Function of the Synaptinemal Complex inLilium Longiflorum Sporocytes". In: *Chromosoma* 23.4, pp. 418–451. DOI: 10.1007/BF00625287 (cit. on p. 17).

Moses, M. J. (Sept. 25, 1958). "The Relation between the Axial Complex of Meiotic Prophase Chromosomes and Chromosome Pairing in a Salamander (Plethodon Cinereus)". In: *The Journal of Biophysical and Biochemical Cytology* 4.5, pp. 633–638. DOI: 10.1083/jcb.4.5.633 (cit. on p. 17).

— (June 1, 1977). "Synaptonemal Complex Karyotyping in Spermatocytes of the Chinese Hamster (Cricetulus Griseus)". In: *Chromosoma* 60.2, pp. 99–125. DOI: 10.1007/BF00288459 (cit. on p. 17).

Gao, J. and M. P. Colaiácovo (Mar. 1, 2018). "Zipping and Unzipping: Protein Modifications Regulating Synaptonemal Complex Dynamics". In: *Trends in Genetics* 34.3, pp. 232–245. DOI: 10.1016/j.tig.2017.12.001 (cit. on p. 17).

Kaniecki, K., L. D. Tullio, and E. C. Greene (Apr. 2018). "A Change of View: Homologous Recombination at Single-Molecule Resolution". In: *Nature Reviews Genetics* 19.4, pp. 191–207. DOI: 10.1038/nrg.2017.92 (cit. on pp. 17, 21).

Dunce, J. M. et al. (July 2018). "Structural Basis of Meiotic Chromosome Synapsis through SYCP1 Self-Assembly". In: *Nature Structural & Molecular Biology* 25.7, pp. 557–569. DOI: 10.1038/s41594-018-0078-9 (cit. on p. 17).

Rankin, S. (July 2015). "Complex Elaboration: Making Sense of Meiotic Cohesin Dynamics". In: *The FEBS journal* 282.13, pp. 2413–2430. DOI: 10.1111/febs.13301. pmid: 25895170 (cit. on pp. 17, 65).

Ishiguro, K.-i. (2019). "The Cohesin Complex in Mammalian Meiosis". In: *Genes to Cells* 24.1, pp. 6–30. DOI: 10.1111/gtc.12652 (cit. on p. 17).

Gaysinskaya, V., B. Miller, et al. (Apr. 4, 2018). *MOESM1 of Transient Reduction of DNA Methylation at the Onset of Meiosis in Male Mice*. DOI: 10.6084/m9.figshare.6094403.v1 (cit. on p. 18).

Gaysinskaya, V., B. F. Miller, et al. (Apr. 4, 2018). "Transient Reduction of DNA Methylation at the Onset of Meiosis in Male Mice". In: *Epigenetics & Chromatin* 11.1, p. 15. DOI: 10.1186/s13072-018-0186-0 (cit. on pp. 18, 114, 119).

Cohen, P. E. and J. K. Holloway (Mar. 1, 2010). "Predicting Gene Networks in Human Oocyte Meiosis". In: *Biology of Reproduction* 82.3, pp. 469–472. DOI: 10.1095/biolreprod.109.083014 (cit. on p. 18).

Bolcun-Filas, E. and J. C. Schimenti (Jan. 1, 2012). "Chapter Five - Genetics of Meiosis and Recombination in Mice". In: *International Review of Cell and Molecular Biology*. Ed. by K. W. Jeon. Vol. 298. Academic Press, pp. 179–227. DOI: 10.1016/B978-0-12-394309-5.00005-5 (cit. on p. 18).

Hughes, S. E. et al. (Mar. 1, 2018). "Female Meiosis: Synapsis, Recombination, and Segregation in Drosophila Melanogaster". In: *Genetics* 208.3, pp. 875–908. DOI: 10.1534/genetics.117.300081. pmid: 29487146 (cit. on p. 18).

Cahoon, C. K. et al. (Aug. 15, 2017). "Superresolution Expansion Microscopy Reveals the Three-Dimensional Organization of the Drosophila Synaptonemal Complex". In: *Proceedings of the National Academy of Sciences* 114.33, E6857–E6866. DOI: 10.1073/pnas.1705623114. pmid: 28760978 (cit. on p. 18).

Sun, H. et al. (Mar. 1989). "Double-Strand Breaks at an Initiation Site for Meiotic Gene Conversion". In: *Nature* 338.6210, pp. 87–90. DOI: 10.1038/338087a0 (cit. on p. 18).

Keeney, S., C. N. Giroux, and N. Kleckner (Feb. 7, 1997). "Meiosis-Specific DNA Double-Strand Breaks Are Catalyzed by Spo11, a Member of a Widely Conserved Protein Family". In: *Cell* 88.3, pp. 375–384. DOI: 10.1016/S0092-8674(00)81876-0 (cit. on p. 18).

Bergerat, A. et al. (Mar. 1997). "An Atypical Topoisomerase II from Archaea with Implications for Meiotic Recombination". In: *Nature* 386.6623, pp. 414–417. DOI: 10.1038/386414a0 (cit. on p. 18).

Cole, F. et al. (Mar. 4, 2012). "Homeostatic Control of Recombination Is Implemented Progressively in Mouse Meiosis". In: *Nature cell biology* 14.4, pp. 424–430. DOI: 10.1038/ncb2451. pmid: 22388890 (cit. on p. 18).

Vrielynck, N. et al. (Feb. 2016). "A DNA Topoisomerase VI-like Complex Initiates Meiotic Recombination". In: *Science* 351.6276, pp. 939–943. DOI: 10.1126/science.aad5196. pmid: 26917763 (cit. on pp. 18, 65).

Robert, T. et al. (Feb. 2016). "The TopoVIB-Like Protein Family Is Required for Meiotic DNA Double-Strand Break Formation". In: *Science* 351.6276, pp. 943–949. DOI: 10.1126/science.aad5309. pmid: 26917764 (cit. on pp. 18, 65).

Li, R. et al. (Aug. 2019). "A High-Resolution Map of Non-Crossover Events Reveals Impacts of Genetic Diversity on Mammalian Meiotic Recombination". In: *Nature communications* 10.1, p. 3900. DOI: 10.1038/s41467-019-11675-y. pmid: 31467277 (cit. on pp. 18, 25, 112, 118).

Neale, M. J., J. Pan, and S. Keeney (Aug. 18, 2005). "Endonucleolytic Processing of Covalent Protein-Linked Double-Strand Breaks". In: *Nature* 436.7053, pp. 1053–1057. DOI: 10.1038/nature03872. pmid: 16107854 (cit. on p. 18).

Kumar, R., H.-M. M. Bourbon, and B. de Massy (June 2010). "Functional Conservation of Mei4 for Meiotic DNA Double-Strand Break Formation from Yeasts to Mice." In: *Genes & development* 24.12, pp. 1266–1280. DOI: 10.1101/gad.571710. pmid: 20551173 (cit. on p. 18).

Panizza, S. et al. (Aug. 5, 2011). "Spo11-Accessory Proteins Link Double-Strand Break Sites to the Chromosome Axis in Early Meiotic Recombination". In: *Cell* 146.3, pp. 372–383. DOI: 10.1016/j.cell.2011.07.003 (cit. on p. 18).

Stanzione, M. et al. (Nov. 2016). "Meiotic DNA Break Formation Requires the Unsynapsed Chromosome Axis-Binding Protein IHO1 (CCDC36) in Mice". In: *Nature cell biology* 18.11, pp. 1208–1220. DOI: 10.1038/ncb3417. pmid: 27723721 (cit. on pp. 18, 65).

Kumar, R. et al. (Dec. 1, 2018). "Mouse REC114 Is Essential for Meiotic DNA Double-Strand Break Formation and Forms a Complex with MEI4". In: *Life Science Alliance* 1.6. DOI: 10.26508/lsa.201800259. pmid: 30569039 (cit. on p. 18).

Papanikos, F. et al. (June 6, 2019). "Mouse ANKRD31 Regulates Spatiotemporal Patterning of Meiotic Recombination Initiation and Ensures Recombination between X and Y Sex Chromosomes". In: *Molecular Cell* 74.5, 1069–1085.e11. DOI: 10.1016/j.molcel.2019.03.022 (cit. on p. 18).

Boekhout, M. et al. (June 6, 2019). "REC114 Partner ANKRD31 Controls Number, Timing, and Location of Meiotic DNA Breaks". In: *Molecular Cell* 74.5, 1053–1068.e8. DOI: 10.1016/j.molcel.2019.03.023 (cit. on pp. 18, 24).

Blat, Y. et al. (Dec. 13, 2002). "Physical and Functional Interactions among Basic Chromosome Organizational Features Govern Early Steps of Meiotic Chiasma Formation". In: *Cell* 111.6, pp. 791–802. DOI: 10.1016/S0092-8674(02)01167-4. pmid: 12526806 (cit. on p. 18).

Acquaviva, L. et al. (Jan. 11, 2013). "The COMPASS Subunit Spp1 Links Histone Methylation to Initiation of Meiotic Recombination". In: *Science* 339.6116, pp. 215–218. DOI: 10.1126/science.1225739. pmid: 23160953 (cit. on p. 19).

Sommermeyer, V. et al. (Jan. 10, 2013). "Spp1, a Member of the Set1 Complex, Promotes Meiotic DSB Formation in Promoters by Tethering Histone H3K4 Methylation Sites to Chromosome Axes". In: *Molecular Cell* 49.1, pp. 43–54. DOI: 10.1016/j.molcel.2012.11.008. pmid: 23246437 (cit. on p. 19).

Tian, H., T. Billings, and P. M. Petkov (Oct. 26, 2018). "CXXC1 Is Not Essential for Normal DNA Double-Strand Break Formation and Meiotic Recombination in Mouse". In: *PLOS Genetics* 14.10, e1007657. DOI: 10.1371/journal.pgen.1007657 (cit. on p. 19).

Kumar, R. and B. De Massy (Dec. 2010). "Initiation of Meiotic Recombination in Mammals". In: *Genes* 1.3, pp. 521–549. DOI: 10.3390/genes1030521 (cit. on p. 19).

Lam, I. and S. Keeney (Jan. 1, 2015a). "Mechanism and Regulation of Meiotic Recombination Initiation". In: *Cold Spring Harbor Perspectives in Biology* 7.1, a016634. DOI: 10.1101/cshperspect.a016634. pmid: 25324213 (cit. on p. 19).

Bhattacharyya, T. et al. (Mar. 18, 2019). "Prdm9 and Meiotic Cohesin Proteins Cooperatively Promote DNA Double-Strand Break Formation in Mammalian Spermatocytes". In: *Current Biology* 29.6, 1002–1018.e7. DOI: 10.1016/j.cub.2019.02.007 (cit. on p. 19).

Daley, J. M. et al. (2014). "Multifaceted Role of the Topo IIIα-RMI1-RMI2 Complex and DNA2 in the BLM-Dependent Pathway of DNA Break End Resection". In: *Nucleic Acids Research* 42.17, pp. 11083–11091. DOI: 10.1093/nar/gku803. pmid: 25200081 (cit. on p. 19).

Symington, L. S. (Jan. 8, 2014). "End Resection at Double-Strand Breaks: Mechanism and Regulation". In: *Cold Spring Harbor Perspectives in Biology* 6.8, a016436. DOI: 10.1101/cshperspect.a016436. pmid: 25085909 (cit. on p. 19).

Schiller, C. B. et al. (Jan. 10, 2014). "Structural Studies of DNA End Detection and Resection in Homologous Recombination". In: *Cold Spring Harbor Perspectives in Biology* 6.10, a017962. DOI: 10.1101/cshperspect.a017962. pmid: 25081516 (cit. on p. 19).

Maréchal, A. and L. Zou (Jan. 9, 2013). "DNA Damage Sensing by the ATM and ATR Kinases". In: *Cold Spring Harbor Perspectives in Biology* 5.9, a012716. DOI: 10.1101/cshperspect.a012716. pmid: 24003211 (cit. on p. 19).

Liu, J. et al. (Oct. 2010). "Human BRCA2 Protein Promotes RAD51 Filament Formation on RPA-Covered ssDNA". In: *Nature structural & molecular biology* 17.10, pp. 1260–1262. DOI: 10.1038/nsmb.1904. pmid: 20729859 (cit. on p. 19).

Jensen, R. B., A. Carreira, and S. C. Kowalczykowski (Oct. 2010). "Purified Human BRCA2 Stimulates RAD51-Mediated Recombination". In: *Nature* 467.7316, pp. 678–683. DOI: 10.1038/nature09399 (cit. on p. 19).

Shi, B. et al. (Feb. 4, 2019). "Dual Functions for the ssDNA-Binding Protein RPA in Meiotic Recombination". In: *PLOS Genetics* 15.2, e1007952. DOI: 10.1371/journal.pgen.1007952 (cit. on p. 19).

Xu, Y. et al. (May 2017). "Meiosis-Specific Proteins MEIOB and SPATA22 Cooperatively Associate with the Single-Stranded DNA-Binding Replication Protein A Complex and DNA Double-Strand Breaks". In: *Biology of reproduction* 96.5, pp. 1096–1104. DOI: 10.1093/biolre/iox040. pmid: 28453612 (cit. on pp. 19, 65).

Sung, P. and H. Klein (Oct. 2006). "Mechanism of Homologous Recombination: Mediators and Helicases Take on Regulatory Functions". In: *Nature Reviews Molecular Cell Biology* 7.10, pp. 739–750. DOI: 10.1038/nrm2008 (cit. on p. 20).

Sehorn, M. G. et al. (May 2004). "Human Meiotic Recombinase Dmc1 Promotes ATP-Dependent Homologous DNA Strand Exchange". In: *Nature* 429.6990, pp. 433–437. DOI: 10.1038/nature02563. pmid: 15164066 (cit. on p. 20).

Cloud, V. et al. (Sept. 7, 2012). "Rad51 Is an Accessory Factor for Dmc1-Mediated Joint Molecule Formation During Meiosis". In: *Science* 337.6099, pp. 1222–1225. DOI: 10.1126/science.1219379. pmid: 22955832 (cit. on p. 20).

Brown, M. S., J. Grubb, et al. (Dec. 2015). "Small Rad51 and Dmc1 Complexes Often Co-Occupy Both Ends of a Meiotic DNA Double Strand Break". In: *PLoS genetics* 11.12, e1005653. DOI: 10.1371/journal.pgen.1005653. pmid: 26719980 (cit. on pp. 20, 65, 104).

Brown, M. S. and D. K. Bishop (Jan. 1, 2015). "DNA Strand Exchange and RecA Homologs in Meiosis". In: *Cold Spring Harbor Perspectives in Biology* 7.1, a016659. DOI: 10.1101/cshperspect.a016659. pmid: 25475089 (cit. on p. 20).

Crickard, J. B. et al. (Mar. 16, 2018). "Spontaneous Self-Segregation of Rad51 and Dmc1 DNA Recombinases within Mixed Recombinase Filaments". In: *Journal of Biological Chemistry* 293.11, pp. 4191–4200. DOI: 10.1074/jbc.RA117.001143. pmid: 29382724 (cit. on p. 20).

Petukhova, G. V. et al. (May 2005). "The Hop2 and Mnd1 Proteins Act in Concert with Rad51 and Dmc1 in Meiotic Recombination". In: *Nature Structural & Molecular Biology* 12.5, pp. 449–453. DOI: 10.1038/nsmb923 (cit. on p. 20).

Gupta, R. C. et al. (Nov. 1, 1999). "Rapid Exchange of A:T Base Pairs Is Essential for Recognition of DNA Homology by Human Rad51 Recombination Protein". In: *Molecular Cell* 4.5, pp. 705–714. DOI: 10.1016/S1097-2765(00)80381-0 (cit. on p. 20).

Folta-Stogniew, E. et al. (Sept. 24, 2004). "Exchange of DNA Base Pairs That Coincides with Recognition of Homology Promoted by E. Coli RecA Protein". In: *Molecular Cell* 15.6, pp. 965–975. DOI: 10.1016/j.molcel.2004.08.017 (cit. on p. 20).

Ragunathan, K., C. Liu, and T. Ha (Dec. 13, 2012). "RecA Filament Sliding on DNA Facilitates Homology Search". In: *eLife* 1. Ed. by X. Zhuang, e00067. DOI: 10.7554/eLife.00067 (cit. on p. 20).

Forget, A. L. and S. C. Kowalczykowski (Feb. 2012). "Single-Molecule Imaging of DNA Pairing by RecA Reveals a Three-Dimensional Homology Search". In: *Nature* 482.7385, pp. 423–427. DOI: 10.1038/nature10782 (cit. on p. 20).

Renkawitz, J. et al. (Apr. 25, 2013). "Monitoring Homology Search during DNA Double-Strand Break Repair In Vivo". In: *Molecular Cell* 50.2, pp. 261–272. DOI: 10.1016/j.molcel.2013.02.020 (cit. on p. 20).

Qi, Z. et al. (Feb. 26, 2015). "DNA Sequence Alignment by Microhomology Sampling during Homologous Recombination". In: *Cell* 160.5, pp. 856–869. DOI: 10.1016/j.cell.2015.01.029 (cit. on p. 20).

Lee, J. Y. et al. (Aug. 28, 2015). "Base Triplet Stepping by the Rad51/RecA Family of Recombinases". In: *Science* 349.6251, pp. 977–981. DOI: 10.1126/science.aab2666. pmid: 26315438 (cit. on p. 20).

Barzel, A. and M. Kupiec (Jan. 2008). "Finding a Match: How Do Homologous Sequences Get Together for Recombination?" In: *Nature Reviews Genetics* 9.1, pp. 27–37. DOI: 10.1038/nrg2224 (cit. on p. 20).

Renkawitz, J., C. A. Lademann, and S. Jentsch (May 2014). "Mechanisms and Principles of Homology Search during Recombination". In: *Nature Reviews Molecular Cell Biology* 15.6, pp. 369–383. DOI: 10.1038/nrm3805 (cit. on p. 20).

Greene, E. C. (May 27, 2016). "DNA Sequence Alignment during Homologous Recombination". In: *Journal of Biological Chemistry* 291.22, pp. 11572–11580. DOI: 10.1074/jbc.R116.724807. pmid: 27129270 (cit. on p. 21).

Haber, J. E. (May 2018). "DNA Repair: The Search for Homology". In: *BioEssays : news and reviews in molecular, cellular and developmental biology* 40.5, e1700229. DOI: 10.1002/bies.201700229. pmid: 29603285 (cit. on p. 21).

Lee, J. Y., Z. Qi, and E. C. Greene (Oct. 14, 2016). "ATP Hydrolysis Promotes Duplex DNA Release by the RecA Presynaptic Complex". In: *Journal of Biological Chemistry* 291.42, pp. 22218–22230. DOI: 10.1074/jbc.M116.740563. pmid: 27587394 (cit. on p. 21).

McNairn, A. J., V. D. Rinaldi, and J. C. Schimenti (Feb. 1, 2017). "Repair of Meiotic DNA Breaks and Homolog Pairing in Mouse Meiosis Requires a Minichromosome Maintenance (MCM) Paralog". In: *Genetics* 205.2, pp. 529–537. DOI: 10.1534/genetics.116.196808. pmid: 27986806 (cit. on p. 21).

McIlwraith, M. J. et al. (Dec. 9, 2005). "Human DNA Polymerase η Promotes DNA Synthesis from Strand Invasion Intermediates of Homologous Recombination". In: *Molecular Cell* 20.5, pp. 783–792. DOI: 10.1016/j.molcel.2005.10.001 (cit. on p. 21).

Kawamoto, T. et al. (Dec. 9, 2005). "Dual Roles for DNA Polymerase η in Homologous DNA Recombination and Translesion DNA Synthesis". In: *Molecular Cell* 20.5, pp. 793–799. DOI: 10.1016/j.molcel.2005.10.016 (cit. on p. 21).

Bugreev, D. V., O. M. Mazina, and A. V. Mazin (Aug. 2006). "Rad54 Protein Promotes Branch Migration of Holliday Junctions". In: *Nature* 442.7102, pp. 590–593. DOI: 10.1038/nature04889 (cit. on p. 21).

Bugreev, D. V., R. M. Brosh, and A. V. Mazin (July 18, 2008). "RECQ1 Possesses DNA Branch Migration Activity". In: *The Journal of Biological Chemistry* 283.29, pp. 20231–20242. DOI: 10.1074/jbc.M801582200. pmid: 18495662 (cit. on p. 21).

Mazina, O. M. et al. (Apr. 6, 2012). "Polarity and Bypass of DNA Heterology during Branch Migration of Holliday Junctions by Human RAD54, BLM, and RECQ1 Proteins". In: *The Journal of Biological Chemistry* 287.15, pp. 11820–11832. DOI: 10.1074/jbc.M112.341347. pmid: 22356911 (cit. on p. 21).

Sugiyama, T., J. H. New, and S. C. Kowalczykowski (May 26, 1998). "DNA Annealing by Rad52 Protein Is Stimulated by Specific Interaction with the Complex of Replication Protein A and Single-Stranded DNA". In: *Proceedings of the National Academy of Sciences* 95.11, pp. 6049–6054. DOI: 10.1073/pnas.95.11.6049. pmid: 9600915 (cit. on p. 21).

Sugiyama, T. et al. (Nov. 29, 2006). "Rad52-Mediated DNA Annealing after Rad51-Mediated DNA Strand Exchange Promotes Second ssDNA Capture". In: *The EMBO Journal* 25.23, pp. 5539–5548. DOI: 10.1038/sj.emboj.7601412 (cit. on p. 21).

McIlwraith, M. J. and S. C. West (Feb. 29, 2008). "DNA Repair Synthesis Facilitates RAD52-Mediated Second-End Capture during DSB Repair". In: *Molecular Cell* 29.4, pp. 510–516. DOI: 10.1016/j.molcel.2007.11.037 (cit. on p. 21).

Lao, J. P. et al. (Feb. 29, 2008). "Rad52 Promotes Postinvasion Steps of Meiotic Double-Strand-Break Repair". In: *Molecular Cell* 29.4, pp. 517–524. DOI: 10.1016/j.molcel.2007.12.014 (cit. on p. 21).

Wu, L. and I. D. Hickson (Dec. 2003). "The Bloom's Syndrome Helicase Suppresses Crossing over during Homologous Recombination". In: *Nature* 426.6968, pp. 870–874. DOI: 10.1038/nature02253 (cit. on p. 21).

Bizard, A. H. and I. D. Hickson (Jan. 7, 2014). "The Dissolution of Double Holliday Junctions". In: *Cold Spring Harbor Perspectives in Biology* 6.7, a016477. DOI: 10.1101/cshperspect.a016477. pmid: 24984776 (cit. on p. 21).

Wyatt, H. D. M. and S. C. West (Jan. 9, 2014). "Holliday Junction Resolvases". In: *Cold Spring Harbor Perspectives in Biology* 6.9, a023192. DOI: 10.1101/cshperspect.a023192. pmid: 25183833 (cit. on p. 21).

Zakharyevich, K. et al. (Apr. 13, 2012). "Delineation of Joint Molecule Resolution Pathways in Meiosis Identifies a Crossover-Specific Resolvase". In: *Cell* 149.2, pp. 334–347. DOI: 10.1016/j.cell.2012.03.023. pmid: 22500800 (cit. on p. 21).

Hunter, N. (Jan. 12, 2015). "Meiotic Recombination: The Essence of Heredity". In: *Cold Spring Harbor Perspectives in Biology* 7.12, a016618. DOI: 10.1101/cshperspect.a016618. pmid: 26511629 (cit. on pp. 22, 26).

Gray, S. and P. E. Cohen (Nov. 23, 2016). "Control of Meiotic Crossovers: From Double-Strand Break Formation to Designation". In: *Annual Review of Genetics* 50.1, pp. 175–210. DOI: 10.1146/annurev-genet-120215-035111 (cit. on p. 22).

Manhart, C. M. and E. Alani (Feb. 1, 2016). "Roles for Mismatch Repair Family Proteins in Promoting Meiotic Crossing Over". In: *DNA Repair*. DNA Mismatch Repair 38, pp. 84–93. DOI: 10.1016/j.dnarep.2015.11.024 (cit. on p. 22).

Toledo, M. et al. (June 6, 2019). "A Mutation in the Endonuclease Domain of Mouse MLH3 Reveals Novel Roles for MutLγ during Crossover Formation in Meiotic Prophase I". In: *PLOS Genetics* 15.6, e1008177. DOI: 10.1371/journal.pgen.1008177 (cit. on p. 22).

Pyatnitskaya, A., V. Borde, and A. De Muyt (Sept. 1, 2019). "Crossing and Zipping: Molecular Duties of the ZMM Proteins in Meiosis". In: *Chromosoma* 128.3, pp. 181–198. DOI: 10.1007/s00412-019-00714-8 (cit. on p. 22).

Rockman, M. V. and L. Kruglyak (Mar. 13, 2009). "Recombinational Landscape and Population Genomics of Caenorhabditis Elegans". In: *PLOS Genetics* 5.3, e1000419. DOI: 10.1371/journal.pgen.1000419 (cit. on p. 22).

Kaur, T. and M. V. Rockman (Jan. 1, 2014). "Crossover Heterogeneity in the Absence of Hotspots in Caenorhabditis Elegans". In: *Genetics* 196.1, pp. 137–148. DOI: 10.1534/genetics.113.158857. pmid: 24172135 (cit. on p. 22).

Chan, A. H., P. A. Jenkins, and Y. S. Song (Dec. 20, 2012). "Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila Melanogaster". In: *PLoS Genetics* 8.12. DOI: 10.1371/journal.pgen.1003090. pmid: 23284288 (cit. on p. 22).

Smukowski Heil, C. S. et al. (Oct. 1, 2015). "Recombining without Hotspots: A Comprehensive Evolutionary Portrait of Recombination in Two Closely Related Species of Drosophila". In: *Genome Biology and Evolution* 7.10, pp. 2829–2842. DOI: 10.1093/gbe/evv182. pmid: 26430062 (cit. on p. 22).

Lam, S. T. et al. (July 1974). "Rec-Mediated Recombinational Hot Spot Activity in Bacteriophage Lambda II. a Mutation Which Causes Hot Spot Activity". In: *Genetics* 77.3, pp. 425–433. pmid: 4415485 (cit. on p. 22).

Myers, R. S. and F. W. Stahl (1994). "χ AND THE RecBC D ENZYME OF ESCHERICHIA COLI". In: *Annual Review of Genetics* 28.1, pp. 49–70. DOI: 10.1146/annurev.ge.28.120194.000405. pmid: 7893137 (cit. on p. 22).

Baudat, F. and A. Nicolas (May 13, 1997). "Clustering of Meiotic Double-Strand Breaks on Yeast Chromosome III". In: *Proceedings of the National Academy of Sciences of the United States of America* 94.10, pp. 5213–5218. pmid: 9144217 (cit. on p. 22).

Wu, T. C. and M. Lichten (Jan. 28, 1994). "Meiosis-Induced Double-Strand Break Sites Determined by Yeast Chromatin Structure". In: *Science* 263.5146, pp. 515–518. DOI: 10.1126/science.8290959. pmid: 8290959 (cit. on p. 22).

Mancera, E. et al. (July 2008). "High-Resolution Mapping of Meiotic Crossovers and Non-Crossovers in Yeast". In: *Nature* 454.7203, pp. 479–485. DOI: 10.1038/nature07135 (cit. on p. 22).

Pan, J. et al. (Mar. 4, 2011). "A Hierarchical Combination of Factors Shapes the Genome-Wide Topography of Yeast Meiotic Recombination Initiation". In: *Cell* 144.5, pp. 719–731. DOI: 10.1016/j.cell.2011.02.009. pmid: 21376234 (cit. on p. 22).

Lam, I. and S. Keeney (Nov. 20, 2015b). "Non-Paradoxical Evolutionary Stability of the Recombination Initiation Landscape in Yeast". In: *Science (New York, N.Y.)* 350.6263, pp. 932–937. DOI: 10.1126/science.aad0814. pmid: 26586758 (cit. on p. 22).

Jeffreys, A. J., L. Kauppi, and R. Neumann (Oct. 2001). "Intensely Punctate Meiotic Recombination in the Class II Region of the Major Histocompatibility Complex". In: *Nature Genetics* 29.2, pp. 217–222. DOI: 10.1038/ng1001-217 (cit. on p. 22).

McVean, G. A. T. et al. (Apr. 23, 2004). "The Fine-Scale Structure of Recombination Rate Variation in the Human Genome". In: *Science (New York, N.Y.)* 304.5670, pp. 581–584. DOI: 10.1126/science.1092500. pmid: 15105499 (cit. on p. 22).

Myers, S. et al. (Oct. 14, 2005). "A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome". In: *Science* 310.5746, pp. 321–324. DOI: 10.1126/science.1117196. pmid: 16224025 (cit. on pp. 22, 25).

The International HapMap Consortium (Oct. 2007). "A Second Generation Human Haplotype Map of over 3.1 Million SNPs". In: *Nature* 449.7164, pp. 851–861. DOI: 10.1038/nature06258 (cit. on p. 22).

Coop, G. et al. (Mar. 7, 2008). "High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans". In: *Science* 319.5868, pp. 1395–1398. DOI: 10.1126/science.1151851. pmid: 18239090 (cit. on p. 22).

Kong, A. et al. (July 2002). "A High-Resolution Recombination Map of the Human Genome". In: *Nature Genetics* 31.3, pp. 241–247. DOI: 10.1038/ng917 (cit. on p. 22).

Kong, A. et al. (Oct. 2010). "Fine-Scale Recombination Rate Differences between Sexes, Populations and Individuals". In: *Nature* 467.7319, pp. 1099–1103. DOI: 10.1038/nature09525 (cit. on p. 22).

Kong, A. et al. (Jan. 2014). "Common and Low-Frequency Variants Associated with Genome-Wide Recombination Rate". In: *Nature genetics* 46.1, pp. 11–16. DOI: 10.1038/ng.2833. pmid: 24270358 (cit. on pp. 22, 66).

Bhérer, C., C. L. Campbell, and A. Auton (Apr. 25, 2017). "Refined Genetic Maps Reveal Sexual Dimorphism in Human Meiotic Recombination at Multiple Scales". In: *Nature Communications* 8. DOI: 10.1038/ncomms14994. pmid: 28440270 (cit. on p. 22).

Halldorsson, B. V. et al. (Jan. 2019). "Characterizing Mutagenic Effects of Recombination through a Sequence-Level Genetic Map". In: *Science* 363.6425. DOI: 10.1126/science.aau1043. pmid: 30679340 (cit. on pp. 22, 66).

Jeffreys, A. J. and R. Neumann (July 2002). "Reciprocal Crossover Asymmetry and Meiotic Drive in a Human Recombination Hot Spot". In: *Nature Genetics* 31.3, pp. 267–271. DOI: 10.1038/ng910 (cit. on p. 22).

Myers, S. et al. (Feb. 2010). "Drive against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination". In: *Science* 327.5967, pp. 876–879. DOI: 10.1126/science.1182363. pmid: 20044541 (cit. on pp. 23, 24).

Baudat, F. et al. (Feb. 12, 2010). "PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice". In: *Science* 327.5967, pp. 836–840. DOI: 10.1126/science.1183439. pmid: 20044539 (cit. on p. 23).

Parvanov, E. D., P. M. Petkov, and K. Paigen (Feb. 12, 2010). "Prdm9 Controls Activation of Mammalian Recombination Hotspots". In: *Science* 327.5967, pp. 835–835. DOI: 10.1126/science.1181495. pmid: 20044538 (cit. on p. 23).

Berg, I. L. et al. (Oct. 2010). "PRDM9 Variation Strongly Influences Recombination Hot-Spot Activity and Meiotic Instability in Humans". In: *Nature Genetics* 42.10, pp. 859–863. DOI: 10.1038/ng.658 (cit. on p. 23).

Powers, N. R. et al. (June 2016). "The Meiotic Recombination Activator PRDM9 Trimethylates Both H3K36 and H3K4 at Recombination Hotspots In Vivo". In: *PLoS genetics* 12.6, e1006146. DOI: 10.1371/journal.pgen.1006146. pmid: 27362481 (cit. on pp. 23, 96).

Diagouraga, B. et al. (Mar. 1, 2018). "PRDM9 Methyltransferase Activity Is Essential for Meiotic DNA Double-Strand Break Formation at Its Binding Sites". In: *Molecular Cell* 69.5, 853–865.e6. DOI: 10.1016/j.molcel.2018.01.033. pmid: 29478809 (cit. on pp. 23, 96).

Birtle, Z. and C. P. Ponting (Dec. 1, 2006). "Meisetz and the Birth of the KRAB Motif". In: *Bioinformatics* 22.23, pp. 2841–2845. DOI: 10.1093/bioinformatics/btl498 (cit. on p. 23).

Imbeault, M., P.-Y. Helleboid, and D. Trono (Mar. 2017). "KRAB Zinc-Finger Proteins Contribute to the Evolution of Gene Regulatory Networks". In: *Nature* 543.7646, pp. 550–554. DOI: 10.1038/nature21683 (cit. on p. 23).

Bruno, M., M. Mahgoub, and T. S. Macfarlan (2019). "The Arms Race Between KRAB–Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals". In: *Annual Review of Genetics* 53.1, pp. 393–416. DOI: 10.1146/annurev-genet-112618-043717. pmid: 31518518 (cit. on p. 23).

Oliver, P. L. et al. (Dec. 4, 2009). "Accelerated Evolution of the Prdm9 Speciation Gene across Diverse Metazoan Taxa". In: *PLOS Genetics* 5.12, e1000753. DOI: 10.1371/journal.pgen.1000753 (cit. on p. 23).

Wall, J. D. et al. (Dec. 1, 2003). "Comparative Linkage-Disequilibrium Analysis of the β-Globin Hotspot in Primates". In: *The American Journal of Human Genetics* 73.6, pp. 1330–1340. DOI: 10.1086/380311 (cit. on p. 23).

Ptak, S. E. et al. (June 15, 2004). "Absence of the TAP2 Human Recombination Hotspot in Chimpanzees". In: *PLOS Biology* 2.6, e155. DOI: 10.1371/journal.pbio.0020155 (cit. on p. 23).

Ptak, S. E. et al. (Apr. 2005). "Fine-Scale Recombination Patterns Differ between Chimpanzees and Humans". In: *Nature Genetics* 37.4, pp. 429–434. DOI: 10.1038/ng1529 (cit. on p. 23).

Winckler, W. et al. (Apr. 1, 2005). "Comparison of Fine-Scale Recombination Rates in Humans and Chimpanzees". In: *Science* 308.5718, pp. 107–111. DOI: 10.1126/science.1105322. pmid: 15705809 (cit. on p. 23).

Auton, A. et al. (Apr. 13, 2012). "A Fine-Scale Chimpanzee Genetic Map from Population Sequencing". In: *Science* 336.6078, pp. 193–198. DOI: 10.1126/science.1216872. pmid: 22422862 (cit. on p. 23).

Paigen, K. et al. (July 11, 2008). "The Recombinational Anatomy of a Mouse Chromosome". In: *PLOS Genetics* 4.7, e1000119. DOI: 10.1371/journal.pgen.1000119 (cit. on p. 23).

Brunschwig, H. et al. (July 2012). "Fine-Scale Maps of Recombination Rates and Hotspots in the Mouse Genome". In: *Genetics* 191.3, pp. 757–764. DOI: 10.1534/genetics.112.141036. pmid: 22562932 (cit. on p. 23).

Booker, T. R., R. W. Ness, and P. D. Keightley (Sept. 2017). "The Recombination Landscape in Wild House Mice Inferred Using Population Genomic Data". In:

Genetics 207.1, pp. 297–309. DOI: 10.1534/genetics.117.300063. pmid: 28751421 (cit. on p. 23).

Smagulova, F. et al. (Apr. 2011). "Genome-Wide Analysis Reveals Novel Molecular Features of Mouse Recombination Hotspots". In: *Nature* 472.7343, pp. 375–378. DOI: 10.1038/nature09869. pmid: 21460839 (cit. on pp. 23, 113).

Khil, P. P. et al. (Jan. 5, 2012). "Sensitive Mapping of Recombination Hotspots Using Sequencing-Based Detection of ssDNA". In: *Genome Research* 22.5, pp. 957–965. DOI: 10.1101/gr.130583.111. pmid: 22367190 (cit. on pp. 23, 109, 112, 113, 128, 129, 156).

Smagulova, F. et al. (Feb. 1, 2016). "The Evolutionary Turnover of Recombination Hot Spots Contributes to Speciation in Mice". In: *Genes & Development* 30.3, pp. 266–280. DOI: 10.1101/gad.270009.115. pmid: 26833728 (cit. on p. 23).

Lange, J. et al. (Oct. 20, 2016). "The Landscape of Mouse Meiotic Double-Strand Break Formation, Processing, and Repair". In: *Cell* 167.3, 695–708.e16. DOI: 10.1016/j.cell.2016.09.035. pmid: 27745971 (cit. on pp. 23, 109, 112, 129, 132).

Hinch, A. G. et al. (Mar. 2019). "Factors Influencing Meiotic Recombination Revealed by Whole-Genome Sequencing of Single Sperm". In: *Science* 363.6433. DOI: 10.1126/science.aau8861. pmid: 30898902 (cit. on pp. 23, 25, 26, 112, 118).

Dréau, A. et al. (Sept. 20, 2019). "Genome-Wide Recombination Map Construction from Single Individuals Using Linked-Read Sequencing". In: *Nature Communications* 10.1, p. 4309. DOI: 10.1038/s41467-019-12210-9. pmid: 31541091 (cit. on p. 23).

Pratto, F. et al. (Nov. 14, 2014a). "Recombination Initiation Maps of Individual Human Genomes". In: *Science* 346.6211. DOI: 10.1126/science.1256442. pmid: 25395542 (cit. on pp. 23, 25).

Bell, A. D. et al. (May 2, 2019). "Insights about Variation in Meiosis from 31,228 Human Sperm Genomes". In: *bioRxiv*, p. 625202. DOI: 10.1101/625202 (cit. on p. 23).

Auton, A. et al. (Dec. 12, 2013). "Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs". In: *PLoS Genetics* 9.12. DOI: 10.1371/journal.pgen.1003984. pmid: 24348265 (cit. on p. 23).

Campbell, C. L. et al. (Sept. 2, 2016). "A Pedigree-Based Map of Recombination in the Domestic Dog Genome". In: *G3: Genes|Genomes|Genetics* 6.11, pp. 3517–3524. DOI: 10.1534/g3.116.034678. pmid: 27591755 (cit. on p. 24).

Muñoz-Fuentes, V., A. D. Rienzo, and C. Vilà (Nov. 10, 2011). "Prdm9, a Major Determinant of Meiotic Recombination Hotspots, Is Not Functional in Dogs and Their Wild Relatives, Wolves and Coyotes". In: *PLOS ONE* 6.11, e25498. DOI: 10.1371/journal.pone.0025498 (cit. on p. 24).

Singhal, S. et al. (Nov. 20, 2015). "Stable Recombination Hotspots in Birds". In: *Science (New York, N.Y.)* 350.6263, pp. 928–932. DOI: 10.1126/science.aad0843. pmid: 26586757 (cit. on p. 24).

Choi, K. et al. (Nov. 2013). "Arabidopsis Meiotic Crossover Hot Spots Overlap with H2A.Z Nucleosomes at Gene Promoters". In: *Nature Genetics* 45.11, pp. 1327–1336. DOI: 10.1038/ng.2766 (cit. on p. 24).

Marand, A. P. et al. (Oct. 30, 2017). "Meiotic Crossovers Are Associated with Open Chromatin and Enriched with Stowaway Transposons in Potato". In: *Genome Biology* 18.1, p. 203. DOI: 10.1186/s13059-017-1326-8 (cit. on p. 24).

Brick, K. et al. (May 2012). "Genetic Recombination Is Directed Away from Functional Genomic Elements in Mice". In: *Nature* 485.7400, pp. 642–645. DOI: 10.1038/nature11089 (cit. on pp. 24, 109, 111, 127, 132).

Hayashi, K., K. Yoshida, and Y. Matsui (Nov. 2005). "A Histone H3 Methyltransferase Controls Epigenetic Events Required for Meiotic Prophase". In: *Nature* 438.7066, pp. 374–378. DOI: 10.1038/nature04112. pmid: 16292313 (cit. on pp. 24, 123).

Mihola, O. et al. (July 2019). "Histone Methyltransferase PRDM9 Is Not Essential for Meiosis in Male Mice". In: *Genome research* 29.7, pp. 1078–1086. DOI: 10.1101/gr.244426.118. pmid: 31186301 (cit. on p. 24).

Narasimhan, V. M. et al. (Apr. 22, 2016). "Health and Population Effects of Rare Gene Knockouts in Adult Humans with Related Parents". In: *Science* 352.6284, pp. 474–477. DOI: 10.1126/science.aac8624. pmid: 26940866 (cit. on p. 24).

Papanikos, F. et al. (2018). "ANKRD31 Regulates Spatiotemporal Patterning of Meiotic Recombination Initiation and Ensures Recombination between Heterologous Sex Chromosomes in Mice". In: DOI: 10.1101/423293 (cit. on pp. 24, 66).

Forejt, J. and P. Iványi (Oct. 1974). "Genetic Studies on Male Sterility of Hybrids between Laboratory and Wild Mice (Mus Musculus L.)" In: *Genetics Research* 24.2, pp. 189–206. DOI: 10.1017/S0016672300015214 (cit. on p. 24).

Mihola, O. et al. (Jan. 16, 2009). "A Mouse Speciation Gene Encodes a Meiotic Histone H3 Methyltransferase". In: *Science* 323.5912, pp. 373–375. DOI: 10.1126/science.1163601. pmid: 19074312 (cit. on p. 24).

Boulton, A., R. S. Myers, and R. J. Redfield (July 22, 1997). "The Hotspot Conversion Paradox and the Evolution of Meiotic Recombination". In: *Proceedings of the National Academy of Sciences of the United States of America* 94.15, pp. 8058–8063. pmid: 9223314 (cit. on p. 24).

Coop, G. and S. R. Myers (Mar. 9, 2007). "Live Hot, Die Young: Transmission Distortion in Recombination Hotspots". In: *PLOS Genetics* 3.3, e35. DOI: 10.1371/journal.pgen.0030035 (cit. on p. 24).

Baker, C. L. et al. (Jan. 8, 2015). "PRDM9 Drives Evolutionary Erosion of Hotspots in Mus Musculus through Haplotype-Specific Initiation of Meiotic Recombination". In: *PLOS Genetics* 11.1, e1004916. DOI: 10.1371/journal.pgen.1004916 (cit. on p. 24).

Altemose, N. et al. (Oct. 2017). "A Map of Human PRDM9 Binding Provides Evidence for Novel Behaviors of PRDM9 and Other Zinc-Finger Proteins in Meiosis". In: *eLife* 6. DOI: 10.7554/eLife.28383. pmid: 29072575 (cit. on pp. 25, 84, 92, 104, 106, 125–127, 129, 132).

Walker, M. et al. (Sept. 7, 2015). "Affinity-Seq Detects Genome-Wide PRDM9 Binding Sites and Reveals the Impact of Prior Chromatin Modifications on Mammalian Recombination Hotspot Usage". In: *Epigenetics & Chromatin* 8.1, p. 31. DOI: 10.1186/s13072-015-0024-6 (cit. on p. 25).

Patel, L. et al. (Mar. 2019). "Dynamic Reorganization of the Genome Shapes the Recombination Landscape in Meiotic Prophase". In: *Nature Structural & Molecular Biology* 26.3 (3), pp. 164–174. DOI: 10.1038/s41594-019-0187-0 (cit. on p. 25).

Yamada, S. et al. (Oct. 18, 2017). "Genomic and Chromatin Features Shaping Meiotic Double-Strand Break Formation and Repair in Mice". In: *Cell Cycle* 16.20, pp. 1870–1884. DOI: 10.1080/15384101.2017.1361065. pmid: 28820351 (cit. on p. 25).

Verver, D. E. et al. (Aug. 2013). "Role for Rodent Smc6 in Pericentromeric Heterochromatin Domains during Spermatogonial Differentiation and Meiosis". In: *Cell Death & Disease* 4.8, e749. DOI: 10.1038/cddis.2013.269. pmid: 23907463 (cit. on p. 25).

Underwood, C. J. et al. (Jan. 4, 2018). "Epigenetic Activation of Meiotic Recombination near Arabidopsis Thaliana Centromeres via Loss of H3K9me2 and Non-CG DNA Methylation". In: *Genome Research* 28.4, pp. 519–531. DOI: 10.1101/gr.227116.117. pmid: 29530927 (cit. on p. 25).

Baudat, F. and B. de Massy (July 1, 2007). "Regulating Double-Stranded DNA Break Repair towards Crossover or Non-Crossover during Mammalian Meiosis". In: *Chromosome Research* 15.5, pp. 565–577. DOI: 10.1007/s10577-007-1140-3 (cit. on p. 25).

Terasawa, M. et al. (Apr. 15, 1995). "Localization of RecA-like Recombination Proteins on Chromosomes of the Lily at Various Meiotic Stages." In: *Genes & Development* 9.8, pp. 925–934. DOI: 10.1101/gad.9.8.925. pmid: 7774810 (cit. on p. 25).

Fledel-Alon, A. et al. (Sept. 18, 2009). "Broad-Scale Recombination Patterns Underlying Proper Disjunction in Humans". In: *PLOS Genetics* 5.9, e1000658. DOI: 10.1371/journal.pgen.1000658 (cit. on p. 26).

Otto, S. P. and B. A. Payseur (2019). "Crossover Interference: Shedding Light on the Evolution of Recombination". In: *Annual Review of Genetics* 53.1, pp. 19–44. DOI: 10.1146/annurev-genet-040119-093957. pmid: 31430178 (cit. on p. 26).

Roeder, G. S. and J. M. Bailis (Sept. 1, 2000). "The Pachytene Checkpoint". In: *Trends in Genetics* 16.9, pp. 395–403. DOI: 10.1016/S0168-9525(00)02080-1 (cit. on p. 26).

Subramanian, V. V. and A. Hochwagen (Jan. 10, 2014). "The Meiotic Checkpoint Network: Step-by-Step through Meiotic Prophase". In: *Cold Spring Harbor Perspectives in Biology* 6.10, a016675. DOI: 10.1101/cshperspect.a016675. pmid: 25274702 (cit. on p. 26).

Watanabe, Y. (June 2012). "Geometry and Force behind Kinetochore Orientation: Lessons from Meiosis". In: *Nature Reviews Molecular Cell Biology* 13.6, pp. 370–382. DOI: 10.1038/nrm3349 (cit. on p. 26).

Buffone, M. G., ed. (2016). *Sperm Acrosome Biogenesis and Function During Fertilization*. Advances in Anatomy, Embryology and Cell Biology. Springer International Publishing. DOI: 10.1007/978-3-319-30567-7 (cit. on p. 26).

Khawar, M. B., H. Gao, and W. Li (2019). "Mechanism of Acrosome Biogenesis in Mammals". In: *Frontiers in Cell and Developmental Biology* 7. DOI: 10.3389/fcell.2019.00195 (cit. on p. 26).

Fawcett, D. W. and D. M. Phillips (1969). "The Fine Structure and Development of the Neck Region of the Mammalian Spermatozoon". In: *The Anatomical Record* 165.2, pp. 153–183. DOI: 10.1002/ar.1091650204 (cit. on p. 26).

Avidor-Reiss, T. and E. L. Fishman (Feb. 1, 2019). "It Takes Two (Centrioles) to Tango". In: *Reproduction* 157.2, R33–R51. DOI: 10.1530/REP-18-0350 (cit. on p. 26).

Linck, R. W., H. Chemes, and D. F. Albertini (Feb. 1, 2016). "The Axoneme: The Propulsive Engine of Spermatozoa and Cilia and Associated Ciliopathies Leading to Infertility". In: *Journal of Assisted Reproduction and Genetics* 33.2, pp. 141–156. DOI: 10.1007/s10815-016-0652-1 (cit. on p. 26).

Lehti, M. S. and A. Sironen (Oct. 1, 2017). "Formation and Function of Sperm Tail Structures in Association with Sperm Motility Defects". In: *Biology of Reproduction* 97.4, pp. 522–536. DOI: 10.1093/biolre/iox096 (cit. on p. 26).

Mitchison, T. J. and H. M. Mitchison (Jan. 2010). "How Cilia Beat". In: *Nature* 463.7279, pp. 308–309. DOI: 10.1038/463308a (cit. on p. 26).

Dunleavy, J. E. M. et al. (Feb. 1, 2019). "The Cytoskeleton in Spermatogenesis". In: *Reproduction* 157.2, R53–R72. DOI: 10.1530/REP-18-0457 (cit. on p. 27).

Wei, Y.-L. and W.-X. Yang (June 20, 2018). "The Acroframosome-Acroplaxome-Manchette Axis May Function in Sperm Head Shaping and Male Fertility". In: *Gene* 660, pp. 28–40. DOI: 10.1016/j.gene.2018.03.059 (cit. on p. 27).

O'Donnell, L. et al. (May 24, 2012). "An Essential Role for Katanin P80 and Microtubule Severing in Male Gamete Production". In: *PLOS Genetics* 8.5, e1002698. DOI: 10.1371/journal.pgen.1002698 (cit. on p. 27).

Kopera, I. A. et al. (May 27, 2010). "Sertoli–Germ Cell Junctions in the Testis: A Review of Recent Data". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1546, pp. 1593–1605. DOI: 10.1098/rstb.2009.0251 (cit. on p. 27).

Wang, W.-L., C.-F. Tu, and Y.-Q. Tan (June 14, 2019). "Insight on Multiple Morphological Abnormalities of Sperm Flagella in Male Infertility: What Is New?" In: (cit. on p. 27).

Gunes, S. et al. (Jan. 1, 2020). "Microtubular Dysfunction and Male Infertility". In: *The World Journal of Men's Health* 38.1, pp. 9–23. DOI: 10.5534/wjmh.180066 (cit. on p. 27).

Kierszenbaum, A. L., E. Rivkin, and L. L. Tres (Aug. 7, 2003). "Acroplaxome, an F-Actin–Keratin-Containing Plate, Anchors the Acrosome to the Nucleus during Shaping of the Spermatid Head". In: *Molecular Biology of the Cell* 14.11, pp. 4628–4640. DOI: 10.1091/mbc.e03-04-0226 (cit. on p. 27).

Kierszenbaum, A. L. and L. L. Tres (2004). "The Acrosome-Acroplaxome-Manchette Complex and the Shaping of the Spermatid Head". In: *Archives of Histology and Cytology* 67.4, pp. 271–284. DOI: 10.1679/aohc.67.271 (cit. on p. 27).

Wong, E. W. P., D. D. Mruk, and C. Y. Cheng (Mar. 1, 2008). "Biology and Regulation of Ectoplasmic Specialization, an Atypical Adherens Junction Type, in the Testis". In: *Biochimica et Biophysica Acta (BBA) - Biomembranes*. Apical Junctional Complexes Part I 1778.3, pp. 692–708. DOI: 10.1016/j.bbamem.2007.11.006 (cit. on p. 27).

Lehti, M. S. and A. Sironen (Apr. 1, 2016). "Formation and Function of the Manchette and Flagellum during Spermatogenesis". In: *Reproduction* 151.4, R43–R54. DOI: 10.1530/REP-15-0310 (cit. on p. 28).

Ho, H.-C. and S. Wey (2007). "Three Dimensional Rendering of the Mitochondrial Sheath Morphogenesis during Mouse Spermiogenesis". In: *Microscopy Research and Technique* 70.8, pp. 719–723. DOI: 10.1002/jemt.20457 (cit. on p. 28).

Toure, A. et al. (2011). "Septins at the Annulus of Mammalian Sperm". In: *Biological Chemistry* 392.8-9, pp. 799–803. DOI: 10.1515/BC.2011.074 (cit. on p. 28).

Kuo, Y.-C. et al. (Mar. 1, 2015). "SEPT12 Orchestrates the Formation of Mammalian Sperm Annulus by Organizing Core Octameric Complexes with Other SEPT Proteins". In: *Journal of Cell Science* 128.5, pp. 923–934. DOI: 10.1242/jcs.158998. pmid: 25588830 (cit. on p. 28).

Balhorn, R. (2007). "The Protamine Family of Sperm Nuclear Proteins". In: *Genome Biology* 8.9, p. 227. DOI: 10.1186/gb-2007-8-9-227. pmid: 17903313 (cit. on p. 28).

Ward, W. S. (Jan. 2010). "Function of Sperm Chromatin Structural Elements in Fertilization and Development". In: *Molecular Human Reproduction* 16.1, pp. 30–36. DOI: 10.1093/molehr/gap080. pmid: 19748904 (cit. on p. 28).

Yamaguchi, K. et al. (June 2018). "Re-Evaluating the Localization of Sperm-Retained Histones Revealed the Modification-Dependent Accumulation in Specific Genome

Regions". In: *Cell Reports* 23.13, pp. 3920–3932. DOI: 10.1016/j.celrep.2018.05.094 (cit. on p. 28).

Cullinane, D. L., T. A. Chowdhury, and K. C. Kleene (Jan. 2015). "Mechanisms of Translational Repression of the Smcp mRNA in Round Spermatids". In: *Reproduction* 149.1, pp. 43–54. DOI: 10.1530/rep-14-0394. pmid: 25336347 (cit. on p. 28).

Kleene, K. C., R. J. Distel, and N. B. Hecht (Sept. 1984). "Translational Regulation and Deadenylation of a Protamine mRNA during Spermiogenesis in the Mouse". In: *Developmental Biology* 105.1, pp. 71–79. DOI: 10.1016/0012-1606(84)90262-8 (cit. on p. 28).

Kleene, K. C. (Mar. 2004). "Patterns, Mechanisms, and Functions of Translation Regulation in Mammalian Spermatogenic Cells". In: *Cytogenetic and Genome Research* 103.3-4, pp. 217–224. DOI: 10.1159/000076807 (cit. on p. 28).

Lacy, D. (Sept. 1, 1962). "CERTAIN ASPECTS OF TESTIS STRUCTURE AND FUNCTION". In: *British Medical Bulletin* 18.3, pp. 205–208. DOI: 10.1093/oxfordjournals.bmb.a069979 (cit. on p. 28).

Breucker, H., E. Schäfer, and A. F. Holstein (1985). "Morphogenesis and Fate of the Residual Body in Human Spermiogenesis". In: *Cell and Tissue Research* 240.2, pp. 303–309. DOI: 10.1007/bf00222339. pmid: 3995554 (cit. on p. 28).

Hermo, L. et al. (Apr. 1, 2010). "Surfing the Wave, Cycle, Life History, and Genes/Proteins Expressed by Testicular Germ Cells. Part 3: Developmental Changes in Spermatid Flagellum and Cytoplasmic Droplet and Interaction of Sperm with the Zona Pellucida and Egg Plasma Membrane". In: *Microscopy Research and Technique* 73.4, pp. 320–363. DOI: 10.1002/jemt.20784 (cit. on p. 28).

Jin, M. et al. (Mar. 22, 2011). "Most Fertilizing Mouse Spermatozoa Begin Their Acrosome Reaction before Contact with the Zona Pellucida during in Vitro Fertilization". In: *Proceedings of the National Academy of Sciences* 108.12, p. 4892. DOI: 10.1073/pnas.1018202108 (cit. on p. 28).

Hikabe, O. et al. (Nov. 2016). "Reconstitution in Vitro of the Entire Cycle of the Mouse Female Germ Line". In: *Nature* 539.7628, pp. 299–303. DOI: 10.1038/nature20104 (cit. on p. 29).

Laiho, A. et al. (Apr. 2013). "Transcriptome Profiling of the Murine Testis during the First Wave of Spermatogenesis". In: *PLoS ONE* 8.4, e61558+. DOI: 10.1371/journal.pone.0061558 (cit. on p. 29).

Ball, R. L. et al. (Aug. 2016). "Regulatory Complexity Revealed by Integrated Cytological and RNA-Seq Analyses of Meiotic Substages in Mouse Spermatocytes". In: *BMC Genomics* 17.1. DOI: 10.1186/s12864-016-2865-1 (cit. on p. 29).

Margolin, G. et al. (2014). "Integrated Transcriptome Analysis of Mouse Spermatogenesis". In: *BMC Genomics* 15.1, pp. 39+. DOI: 10.1186/1471-2164-15-39 (cit. on p. 29).

Li, Y., D. Ray, and P. Ye (Feb. 2013). "Identification of Germ Cell-Specific Genes in Mammalian Meiotic Prophase". In: *BMC Bioinformatics* 14.1, pp. 72+. DOI: 10.1186/1471-2105-14-72 (cit. on p. 29).

Soh, Y. Q. S. et al. (Sept. 2015). "A Gene Regulatory Program for Meiotic Prophase in the Fetal Ovary". In: *PLoS Genet* 11.9, e1005531+. DOI: 10.1371/journal.pgen.1005531 (cit. on pp. 29, 96, 140).

Soumillon, M. et al. (June 2013). "Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis". In: *Cell Reports* 3.6, pp. 2179–2190. DOI: 10.1016/j.celrep.2013.05.031 (cit. on pp. 29, 68, 138).

Buard, J. et al. (Sept. 2009). "Distinct Histone Modifications Define Initiation and Repair of Meiotic Recombination in the Mouse." In: *The EMBO journal* 28.17, pp. 2616–2624. DOI: 10.1038/emboj.2009.207. pmid: 19644444 (cit. on p. 29).

Grabske, R. J. et al. (Aug. 1975). "Centrifugal Elutriation: Separation of Spermatogenic Cells on the Basis of Sedimentation Velocity". In: *J. Cell. Physiol.* 86.1, pp. 177–189. DOI: 10.1002/jcp.1040860119 (cit. on p. 29).

Da Cruz, I. et al. (Apr. 2016). "Transcriptome Analysis of Highly Purified Mouse Spermatogenic Cell Populations: Gene Expression Signatures Switch from Meiotic-to Postmeiotic-Related Processes at Pachytene Stage". In: *BMC genomics* 17, p. 294. DOI: 10.1186/s12864-016-2618-1. pmid: 27094866 (cit. on pp. 29, 68, 69).

Brar, G. A. et al. (Dec. 2011). "High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling". In: *Science* 335.6068, pp. 552–557. DOI: 10.1126/science.1215110. pmid: 22194413 (cit. on p. 29).

Mata, J. et al. (Aug. 2002). "The Transcriptional Program of Meiosis and Sporulation in Fission Yeast". In: *Nature Genetics* 32.1, pp. 143–147. DOI: 10.1038/ng951. pmid: 12161753 (cit. on p. 29).

Chu, S. et al. (Oct. 1998). "The Transcriptional Program of Sporulation in Budding Yeast". In: *Science* 282.5389, pp. 699–705. DOI: 10.1126/science.282.5389.699. pmid: 9784122 (cit. on p. 29).

Handel, M. A. and J. C. Schimenti (Jan. 2010). "Genetics of Mammalian Meiosis: Regulation, Dynamics and Impact on Fertility". In: *Nature Reviews Genetics* 11.2, pp. 124–136. DOI: 10.1038/nrg2723 (cit. on p. 29).

Zappia, L., B. Phipson, and A. Oshlack (June 25, 2018). "Exploring the Single-Cell RNA-Seq Analysis Landscape with the scRNA-Tools Database". In: *PLOS Computational Biology* 14.6, e1006245. DOI: 10.1371/journal.pcbi.1006245 (cit. on p. 30).

Žurauskienė, J. and C. Yau (Mar. 2016). "pcaReduce: Hierarchical Clustering of Single Cell Transcriptional Profiles". In: *BMC Bioinformatics* 17.1. DOI: 10.1186/s12859-016-0984-y (cit. on p. 30).

Kiselev, V. Y. et al. (Mar. 2017). "SC3: Consensus Clustering of Single-Cell RNA-Seq Data". In: *Nature Methods* 14.5, pp. 483–486. DOI: 10.1038/nmeth.4236 (cit. on p. 30).

Guo, M. et al. (Nov. 2015). "SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis". In: *PLOS Computational Biology* 11.11, e1004575+. DOI: 10.1371/journal.pcbi.1004575 (cit. on p. 30).

Satija, R. et al. (Apr. 2015). "Spatial Reconstruction of Single-Cell Gene Expression Data". In: *Nature Biotechnology* 33.5, pp. 495–502. DOI: 10.1038/nbt.3192 (cit. on p. 30).

Trapnell, C. et al. (Mar. 2014). "The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells". In: *Nature Biotechnology* 32.4, pp. 381–386. DOI: 10.1038/nbt.2859. pmid: 24658644 (cit. on p. 30).

Ji, Z. and H. Ji (July 2016). "TSCAN: Pseudo-Time Reconstruction and Evaluation in Single-Cell RNA-Seq Analysis." In: *Nucleic acids research* 44.13, e117. DOI: 10.1093/nar/gkw430. pmid: 27179027 (cit. on p. 30).

Marco, E. et al. (Dec. 2014). "Bifurcation Analysis of Single-Cell Gene Expression Data Reveals Epigenetic Landscape". In: *Proceedings of the National Academy of Sciences of the United States of America* 111.52, E5643–50. DOI: 10.1073/pnas.1408993111. pmid: 25512504 (cit. on pp. 30, 45).

Qiu, X. et al. (Feb. 2017). "Reversed Graph Embedding Resolves Complex Single-Cell Developmental Trajectories". In: *bioRxiv*, pp. 110668+. DOI: 10.1101/110668 (cit. on p. 30).

Donoho, D. L. (2000). "High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality". In: p. 32 (cit. on p. 30).

Moon, K. R. et al. (Feb. 1, 2018). "Manifold Learning-Based Methods for Analyzing Single-Cell RNA-Sequencing Data". In: *Current Opinion in Systems Biology.* • Future of Systems Biology • Genomics and Epigenomics 7, pp. 36–46. DOI: 10.1016/j.coisb.2017.12.008 (cit. on p. 30).

McInnes, L., J. Healy, N. Saul, et al. (2018). "UMAP: Uniform Manifold Approximation and Projection". In: *Journal of Open Source Software* 3.29, p. 861. DOI: 10.21105/joss.00861 (cit. on p. 31).

Becht, E. et al. (Jan. 2019). "Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP". In: *Nature Biotechnology* 37.1, pp. 38–44. DOI: 10.1038/nbt.4314 (cit. on p. 31).

Haghverdi, L., F. Buettner, and F. J. Theis (Sept. 2015). "Diffusion Maps for High-Dimensional Single-Cell Analysis of Differentiation Data". In: *Bioinformatics* 31.18, pp. 2989–2998. DOI: 10.1093/bioinformatics/btv325. pmid: 26002886 (cit. on p. 31).

Moon, K. R. et al. (Dec. 2019). "Visualizing Structure and Transitions in High-Dimensional Biological Data". In: *Nature Biotechnology* 37.12, pp. 1482–1492. DOI: 10.1038/s41587-019-0336-3 (cit. on p. 31).

Farrell, J. A. et al. (June 1, 2018). "Single-Cell Reconstruction of Developmental Trajectories during Zebrafish Embryogenesis". In: *Science* 360.6392. DOI: 10.1126/science.aar3131. pmid: 29700225 (cit. on p. 31).

Weinreb, C., S. Wolock, and A. M. Klein (Apr. 1, 2018). "SPRING: A Kinetic Interface for Visualizing High Dimensional Single-Cell Expression Data". In: *Bioinformatics* 34.7, pp. 1246–1248. DOI: 10.1093/bioinformatics/btx792 (cit. on p. 31).

Wagner, D. E. et al. (June 1, 2018). "Single-Cell Mapping of Gene Expression Landscapes and Lineage in the Zebrafish Embryo". In: *Science* 360.6392, pp. 981–987. DOI: 10.1126/science.aar4362. pmid: 29700229 (cit. on p. 31).

Welch, J. D., A. J. Hartemink, and J. F. Prins (May 2016). "SLICER: Inferring Branched, Nonlinear Cellular Trajectories from Single Cell RNA-Seq Data". In: *Genome Biology* 17.1. DOI: 10.1186/s13059-016-0975-3 (cit. on p. 31).

Campbell, K. and C. Yau (Sept. 2015). "Bayesian Gaussian Process Latent Variable Models for Pseudotime Inference in Single-Cell RNA-Seq Data". In: *bioRxiv*, pp. 026872+. DOI: 10.1101/026872 (cit. on p. 31).

Alter, O., P. O. Brown, and D. Botstein (Aug. 2000). "Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling". In: *Proceedings of the National Academy of Sciences of the United States of America* 97.18, pp. 10101–10106. DOI: 10.1073/pnas.97.18.10101. pmid: 10963673 (cit. on p. 31).

Green, C. D. et al. (Sept. 2018). "A Comprehensive Roadmap of Murine Spermatogenesis Defined by Single-Cell RNA-Seq". In: *Developmental cell* 46.5, 651–667.e10. DOI: 10.1016/j.devcel.2018.07.025. pmid: 30146481 (cit. on pp. 31, 62, 134, 136).

Brunet, J.-P. et al. (Mar. 2004). "Metagenes and Molecular Pattern Discovery Using Matrix Factorization". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.12, pp. 4164–4169. DOI: 10.1073/pnas.0308531101. pmid: 15016911 (cit. on p. 31).

Kim, P. M. and B. Tidor (July 2003). "Subsystem Identification through Dimensionality Reduction of Large-Scale Gene Expression Data". In: *Genome research* 13.7, pp. 1706–1718. DOI: 10.1101/gr.903503. pmid: 12840046 (cit. on p. 31).

Shao, C. and T. Höfer (Jan. 2017). "Robust Classification of Single-Cell Transcriptome Data by Nonnegative Matrix Factorization". In: *Bioinformatics* 33.2, pp. 235–242. DOI: 10.1093/bioinformatics/btw607. pmid: 27663498 (cit. on p. 31).

Zhu, X. et al. (Jan. 2017). "Detecting Heterogeneity in Single-Cell RNA-Seq Data by Non-Negative Matrix Factorization". In: *PeerJ* 5, e2888. DOI: 10.7717/peerj.2888. pmid: 28133571 (cit. on p. 31).

Duren, Z. et al. (July 2018). "Integrative Analysis of Single-Cell Genomics Data by Coupled Nonnegative Matrix Factorizations". In: *Proceedings of the National Academy of Sciences of the United States of America* 115.30, pp. 7723–7728. DOI: 10.1073/pnas.1805681115. pmid: 29987051 (cit. on p. 31).

Kotliar, D. et al. (Apr. 2018). "Identifying Gene Expression Programs of Cell-Type Identity and Cellular Activity with Single-Cell RNA-Seq". In: DOI: 10.7554/eLife.43803 (cit. on p. 31).

Welch, J. D. et al. (June 13, 2019). "Single-Cell Multi-Omic Integration Compares and Contrasts Features of Brain Cell Identity". In: *Cell* 177.7, 1873–1887.e17. DOI: 10.1016/j.cell.2019.05.006. pmid: 31178122 (cit. on p. 31).

Bernardo, J. M. (July 2003). *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*. Oxford University Press (cit. on p. 31).

Buettner, F. et al. (Jan. 2015). "Computational Analysis of Cell-to-Cell Heterogeneity in Single-Cell RNA-Sequencing Data Reveals Hidden Subpopulations of Cells". In: *Nature Biotechnology* 33.2, pp. 155–160. DOI: 10.1038/nbt.3102 (cit. on p. 31).

Stein-O'Brien, G. L. et al. (Oct. 2018). "Enter the Matrix: Factorization Uncovers Knowledge from Omics". In: *Trends in genetics: TIG* 34.10, pp. 790–805. DOI: 10.1016/j.tig.2018.07.003. pmid: 30143323 (cit. on p. 31).

Pierson, E. and C. Yau (Nov. 2015). "ZIFA: Dimensionality Reduction for Zero-Inflated Single-Cell Gene Expression Analysis". In: *Genome Biology* 16.1, pp. 241+. DOI: 10.1186/s13059-015-0805-z (cit. on p. 32).

Vieth, B. et al. (Nov. 1, 2017). "powsimR: Power Analysis for Bulk and Single Cell RNA-Seq Experiments". In: *Bioinformatics* 33.21, pp. 3486–3488. DOI: 10.1093/bioinformatics/btx435 (cit. on p. 32).

Svensson, V. (Mar. 19, 2019). "Droplet scRNA-Seq Is Not Zero-Inflated". In: *bioRxiv*, p. 582064. DOI: 10.1101/582064 (cit. on p. 32).

Townes, F. W. et al. (Mar. 11, 2019). "Feature Selection and Dimension Reduction for Single Cell RNA-Seq Based on a Multinomial Model". In: *bioRxiv*, p. 574574. DOI: 10.1101/574574 (cit. on p. 32).

Löffler-Wirth, H., M. Kalcher, and H. Binder (Oct. 2015). "oposSOM: R-Package for High-Dimensional Portraying of Genome-Wide Expression Landscapes on Bioconductor". In: *Bioinformatics* 31.19, pp. 3225–3227. DOI: 10.1093/bioinformatics/btv342. pmid: 26063839 (cit. on p. 32).

Kim, D. H. et al. (Jan. 2015). "Single-Cell Transcriptome Analysis Reveals Dynamic Changes in lncRNA Expression during Reprogramming". In: *Cell Stem Cell* 16.1, pp. 88–101. DOI: 10.1016/j.stem.2014.11.005 (cit. on p. 32).

Eraslan, G. et al. (Jan. 2019). "Single-Cell RNA-Seq Denoising Using a Deep Count Autoencoder". In: *Nature communications* 10.1, p. 390. DOI: 10.1038/s41467-018-07931-2. pmid: 30674886 (cit. on p. 32).

Lopez, R. et al. (Dec. 2018). "Deep Generative Modeling for Single-Cell Transcriptomics". In: *Nature Methods* 15.12, pp. 1053–1058. DOI: 10.1038/s41592-018-0229-2 (cit. on p. 32).

Svensson, V. and L. Pachter (Aug. 16, 2019). "Interpretable Factor Models of Single-Cell RNA-Seq via Variational Autoencoders". In: *bioRxiv*, p. 737601. DOI: 10.1101/737601 (cit. on p. 33).

Hore, V. and J. Marchini (2015). "Latent Variable Models for Analysing Multidimensional Gene Expression Data". PhD Thesis. The University of Oxford (cit. on pp. 33, 35, 36, 139).

Hore, V. et al. (Sept. 2016). "Tensor Decomposition for Multiple-Tissue Gene Expression Experiments". In: *Nature genetics* 48.9, pp. 1094–1100. DOI: 10.1038/ng.3624. pmid: 27479908 (cit. on pp. 33, 35, 36, 42).

Jaakkola, T. S. and M. I. Jordan (Jan. 1, 2000). "Bayesian Parameter Estimation via Variational Methods". In: *Statistics and Computing* 10.1, pp. 25–37. DOI: 10.1023/A:1008932416310 (cit. on p. 35).

West, M. (2003). "Bayesian Factor Regression Models in the "Large p, Small n" Paradigm". In: *Bayesian Statistics*. Oxford University Press, pp. 723–732 (cit. on p. 36).

Lucas, J. et al. (July 2006). *Sparse Statistical Modelling in Gene Expression Genomics*. URL: /core/books/bayesian-inference-for-gene-expression-and-proteomics/sparse-statistical-modelling-in-gene-expression-genomics/B7F8280805654F38E6E7A9D31EF650EB (visited on 12/21/2019) (cit. on p. 36).

Schwann, T. (1839). *Mikroskopische Untersuchungen Über Die Uebereinstimmung in Der Struktur Und Dem Wachsthum Der Thiere Und Pflanzen*. Sander (cit. on p. 37).

Jung, M. et al. (June 25, 2019). "Unified Single-Cell Analysis of Testis Gene Regulation and Pathology in Five Mouse Strains". In: *eLife* 8. Ed. by D. Bourc'his, P. J. Wittkopp, and S. Lukassen, e43966. DOI: 10.7554/eLife.43966 (cit. on pp. 38, 39, 90, 100, 116, 119, 134).

Lima, A. C. et al. (July 2017). "A Standardized Approach for Multispecies Purification of Mammalian Male Germ Cells by Mechanical Tissue Dissociation and Flow Cytometry". In: *Journal of visualized experiments: JoVE* 125. DOI: 10.3791/55913. pmid: 28745623 (cit. on p. 39).

Szabó, P. E. et al. (July 1, 2002). "Allele-Specific Expression of Imprinted Genes in Mouse Migratory Primordial Germ Cells". In: *Mechanisms of Development* 115.1, pp. 157–160. DOI: 10.1016/S0925-4773(02)00087-4 (cit. on p. 39).

Lipkin, S. M. et al. (Aug. 2002). "Meiotic Arrest and Aneuploidy in MLH3-Deficient Mice". In: *Nature genetics* 31.4, pp. 385–390. DOI: 10.1038/ng931. pmid: 12091911 (cit. on pp. 39, 74).

Shin, Y.-H. et al. (Nov. 2010). "Hormad1 Mutation Disrupts Synaptonemal Complex Formation, Recombination, and Chromosome Segregation in Mammalian Meiosis". In: *PLoS genetics* 6.11, e1001190. DOI: 10.1371/journal.pgen.1001190. pmid: 21079677 (cit. on pp. 39, 71).

Yin, Y. et al. (Aug. 2011). "The E3 Ubiquitin Ligase Cullin 4A Regulates Meiotic Progression in Mouse Spermatogenesis". In: *Developmental biology* 356.1, pp. 51–62. DOI: 10.1016/j.ydbio.2011.05.661. pmid: 21624359 (cit. on pp. 39, 74).

Ilicic, T. et al. (Feb. 2016). "Classification of Low Quality Cells from Single-Cell RNA-Seq Data". In: *Genome biology* 17, p. 29. DOI: 10.1186/s13059-016-0888-1. pmid: 26887813 (cit. on p. 40).

Maaten, L. van der and G. Hinton (2008). "Visualizing Data Using T-SNE". In: *Journal of machine learning research: JMLR* 9 (Nov), pp. 2579–2605 (cit. on p. 45).

McInnes, L., J. Healy, and J. Melville (2018). "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: (cit. on p. 45).

Becht, E. et al. (Dec. 3, 2018). "Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP". In: *Nature Biotechnology* 37 (cit. on p. 45).

Novembre, J. and M. Stephens (Apr. 20, 2008). "Interpreting Principal Component Analyses of Spatial Population Genetic Variation". In: *Nature Genetics* 40 (cit. on p. 45).

Podani, J. and I. Miklos (2002). "RESEMBLANCE COEFFICIENTS AND THE HORSESHOE EFFECT IN PRINCIPAL COORDINATES ANALYSIS". In: *Ecology* 83.12, pp. 3331–3343. DOI: 10.1890/0012-9658(2002)083[3331:RCATHE]2.0.CO;2 (cit. on p. 45).

Hastie, T. and W. Stuetzle (June 1989). "Principal Curves". In: *Journal of the American Statistical Association* 84.406, pp. 502–516. DOI: 10.1080/01621459.1989.10478797 (cit. on p. 45).

Van Dijk, D. et al. (July 2018). "Recovering Gene Interactions from Single-Cell Data Using Data Diffusion". In: *Cell* 174.3, 716–729.e27. DOI: 10.1016/j.cell.2018.05.061. pmid: 29961576 (cit. on pp. 51, 53, 91, 92).

Marini, M. et al. (Oct. 2018). "Reappraising the Microscopic Anatomy of Human Testis: Identification of Telocyte Networks in the Peritubular and Intertubular Stromal Space". In: *Scientific reports* 8.1, p. 14780. DOI: 10.1038/s41598-018-33126-2. pmid: 30283023 (cit. on pp. 62, 136).

Kuroda, N. et al. (July 2004). "Distribution and Role of CD34-Positive Stromal Cells and Myofibroblasts in Human Normal Testicular Stroma". In: *Histology and histopathology* 19.3, pp. 743–751. DOI: 10.14670/HH-19.743. pmid: 15168336 (cit. on p. 62).

Stojkov, N. J. et al. (Mar. 2013). "Orally Applied Doxazosin Disturbed Testosterone Homeostasis and Changed the Transcriptional Profile of Steroidogenic Machinery, cAMP/cGMP Signalling and Adrenergic Receptors in Leydig Cells of Adult Rats". In: *Andrology* 1.2, pp. 332–347. DOI: 10.1111/j.2047-2927.2012.00035.x (cit. on pp. 62, 73).

Balvers, M. et al. (June 1, 1998). "Relaxin-Like Factor Expression as a Marker of Differentiation in the Mouse Testis and Ovary". In: *Endocrinology* 139.6, pp. 2960–2970. DOI: 10.1210/endo.139.6.6046 (cit. on p. 62).

Baker, P. J. and P. J. O'Shaughnessy (Oct. 1, 2001). "Expression of Prostaglandin D Synthetase during Development in the Mouse Testis". In: *Reproduction* 122.4, pp. 553–559. DOI: 10.1530/rep.0.1220553 (cit. on p. 62).

Sanz, E. et al. (June 2013). "RiboTag Analysis of Actively Translated mRNAs in Sertoli and Leydig Cells In Vivo". In: *PLOS ONE* 8.6, e66179+. DOI: 10.1371/journal.pone.0066179 (cit. on p. 63).

Matsui, H., A. Moriyama, and T. Takahashi (2000). "Cloning and Characterization of Mouse Klk27, a Novel Tissue Kallikrein Expressed in Testicular Leydig Cells and Exhibiting Chymotrypsin-like Specificity". In: *European Journal of Biochemistry* 267.23, pp. 6858–6865. DOI: 10.1046/j.1432-1033.2000.01786.x (cit. on p. 63).

Matsui, H. and T. Takahashi (Nov. 1, 2001). "Mouse Testicular Leydig Cells Express Klk21, a Tissue Kallikrein That Cleaves Fibronectin and IGF-Binding Protein-3". In: *Endocrinology* 142.11, pp. 4918–4929. DOI: 10.1210/endo.142.11.8505 (cit. on p. 63).

Matsui, H., N. Takano, and T. Takahashi (Nov. 1, 2005). "Characterization of Mouse Glandular Kallikrein 24 Expressed in Testicular Leydig Cells". In: *The International Journal of Biochemistry & Cell Biology* 37.11, pp. 2333–2343. DOI: 10.1016/j.biocel.2005.05.011 (cit. on p. 63).

Suetsugu, S. et al. (2007). "Male-Specific Sterility Caused by the Loss of CR16". In: *Genes to Cells* 12.6, pp. 721–733. DOI: 10.1111/j.1365-2443.2007.01088.x (cit. on p. 63).

Gehin, M. et al. (Aug. 15, 2002). "The Function of TIF2/GRIP1 in Mouse Reproduction Is Distinct from Those of SRC-1 and p/CIP". In: *Molecular and Cellular Biology* 22.16, pp. 5923–5937. DOI: 10.1128/MCB.22.16.5923-5937.2002. pmid: 12138202 (cit. on p. 63).

Zhou, J. et al. (2011). "Nxf3 Is Expressed in Sertoli Cells, but Is Dispensable for Spermatogenesis". In: *Molecular Reproduction and Development* 78.4, pp. 241–249. DOI: 10.1002/mrd.21291 (cit. on p. 63).

Mossadegh-Keller, N. et al. (Oct. 2, 2017). "Developmental Origin and Maintenance of Distinct Testicular Macrophage Populations". In: *Journal of Experimental Medicine* 214.10, pp. 2829–2841. DOI: 10.1084/jem.20170829. pmid: 28784628 (cit. on p. 63).

Fabriek, B. O., C. D. Dijkstra, and T. K. van den Berg (Aug. 19, 2005). "The Macrophage Scavenger Receptor CD163". In: *Immunobiology* 210.2, pp. 153–160. DOI: 10.1016/j.imbio.2005.05.010 (cit. on p. 63).

Sasmono, R. T. and E. Williams (2012). "Generation and Characterization of MacGreen Mice, the Cfs1r-EGFP Transgenic Mice". In: *Leucocytes: Methods and Protocols*. Ed. by R. B. Ashman. Methods in Molecular Biology. Totowa, NJ: Humana Press, pp. 157–176. DOI: 10.1007/978-1-61779-527-5_11 (cit. on p. 63).

Frungieri, M. et al. (Aug. 2002). "Number, Distribution Pattern, and Identification of Macrophages in the Testes of Infertile Men". In: *Fertility and sterility* 78.2, pp. 298–306. DOI: 10.1016/S0015-0282(02)03206-5 (cit. on p. 63).

Goluža, T. et al. (May 2014). "Macrophages and Leydig Cells in Testicular Biopsies of Azoospermic Men". In: *BioMed research international* 2014, p. 828697. DOI: 10.1155/2014/828697. pmid: 24895614 (cit. on p. 63).

Murray, P. J. and T. A. Wynn (Nov. 2011). "Protective and Pathogenic Functions of Macrophage Subsets". In: *Nature Reviews Immunology* 11.11, pp. 723–737. DOI: 10.1038/nri3073 (cit. on p. 64).

Murphy, K. et al. (2012). *Janeway's Immunobiology*. New York: Garland Science (cit. on p. 64).

Xu, H. et al. (Nov. 2010). "MS4a4B, a CD20 Homologue in T Cells, Inhibits T Cell Propagation by Modulation of Cell Cycle". In: *PLOS ONE* 5.11, e13780. DOI: 10.1371/journal.pone.0013780 (cit. on p. 64).

Call, M. E. et al. (Dec. 27, 2002). "The Organizing Principle in the Formation of the T Cell Receptor-CD3 Complex". In: *Cell* 111.7, pp. 967–979. DOI: 10.1016/S0092-8674(02)01194-7 (cit. on p. 64).

Mayerhofer, A. (May 1, 2013). "Human Testicular Peritubular Cells: More than Meets the Eye". In: *Reproduction* 145.5, R107–R116. DOI: 10.1530/REP-12-0497 (cit. on p. 64).

Buaas, F. W. et al. (June 2004). "Plzf Is Required in Adult Male Germ Cells for Stem Cell Self-Renewal". In: *Nature genetics* 36.6, pp. 647–652. DOI: 10.1038/ng1366. pmid: 15156142 (cit. on p. 64).

Goertz, M. J. et al. (Sept. 2011). "Foxo1 Is Required in Mouse Spermatogonial Stem Cells for Their Maintenance and the Initiation of Spermatogenesis". In: *The Journal of clinical investigation* 121.9, pp. 3456–3466. DOI: 10.1172/JCI57984. pmid: 21865646 (cit. on p. 64).

Kang, H. S. et al. (Nov. 2016). "Transcription Factor GLIS3: A New and Critical Regulator of Postnatal Stages of Mouse Spermatogenesis". In: *Stem cells* 34.11, pp. 2772–2783. DOI: 10.1002/stem.2449. pmid: 27350140 (cit. on p. 64).

Meng, X. et al. (Feb. 25, 2000). "Regulation of Cell Fate Decision of Undifferentiated Spermatogonia by GDNF". In: *Science* 287.5457, pp. 1489–1493. DOI: 10.1126/science.287.5457.1489. pmid: 10688798 (cit. on p. 65).

He, Z. et al. (2007). "Gfra1 Silencing in Mouse Spermatogonial Stem Cells Results in Their Differentiation Via the Inactivation of RET Tyrosine Kinase1". In: *Biology of reproduction* 77.4, pp. 723–733. DOI: 10.1095/biolreprod.107.062513 (cit. on p. 65).

Suzuki, H. et al. (Dec. 2009). "The Heterogeneity of Spermatogonia Is Revealed by Their Topology and Expression of Marker Proteins Including the Germ Cell-Specific Proteins Nanos2 and Nanos3". In: *Developmental biology* 336.2, pp. 222–231. DOI: 10.1016/j.ydbio.2009.10.002. pmid: 19818747 (cit. on p. 65).

Zheng, K. et al. (June 2009). "The Pluripotency Factor LIN28 Marks Undifferentiated Spermatogonia in Mouse". In: *BMC developmental biology* 9, p. 38. DOI: 10.1186/1471-213X-9-38. pmid: 19563657 (cit. on p. 65).

Manova, K. et al. (Dec. 1, 1990). "Gonadal Expression of C-Kit Encoded at the W Locus of the Mouse". In: *Development* 110.4, pp. 1057–1069. pmid: 1712701 (cit. on p. 65).

Schrans-Stassen, B. H. G. J. et al. (Dec. 1, 1999). "Differential Expression of C-Kit in Mouse Undifferentiated and Differentiating Type A Spermatogonia". In: *Endocrinology* 140.12, pp. 5894–5900. DOI: 10.1210/endo.140.12.7172 (cit. on p. 65).

Yoshinaga, K. et al. (Oct. 1, 1991). "Role of C-Kit in Mouse Spermatogenesis: Identification of Spermatogonia as a Specific Site of c-Kit Expression and Function". In: *Development* 113.2, pp. 689–699. pmid: 1723681 (cit. on p. 65).

Oulad-Abdelghani, M. et al. (Oct. 15, 1996). "Characterization of a Premeiotic Germ Cell-Specific Cytoplasmic Protein Encoded by Stra8, a Novel Retinoic Acid-Responsive Gene." In: *The Journal of Cell Biology* 135.2, pp. 469–477. DOI: 10.1083/jcb.135.2.469. pmid: 8896602 (cit. on p. 65).

Endo, T. et al. (May 2015). "Periodic Retinoic Acid–STRA8 Signaling Intersects with Periodic Germ-Cell Competencies to Regulate Spermatogenesis". In: *Proceedings of the National Academy of Sciences* 112.18, E2347–E2356. DOI: 10.1073/pnas.1505683112. pmid: 25902548 (cit. on p. 65).

Baltus, A. E. et al. (Dec. 2006). "In Germ Cells of Mouse Embryonic Ovaries, the Decision to Enter Meiosis Precedes Premeiotic DNA Replication". In: *Nature Genetics* 38.12, pp. 1430–1434. DOI: 10.1038/ng1919 (cit. on p. 65).

Sleutels, F. et al. (June 18, 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". In: *Epigenetics & Chromatin* 5.1, p. 8. DOI: 10.1186/1756-8935-5-8 (cit. on p. 65).

Budhram-Mahadeo, V. et al. (Oct. 1, 2001). "The Closely Related POU Family Transcription Factors Brn-3a and Brn-3b Are Expressed in Distinct Cell Types in the Testis". In: *The International Journal of Biochemistry & Cell Biology* 33.10, pp. 1027–1039. DOI: 10.1016/S1357-2725(01)00069-3 (cit. on p. 65).

Maezawa, S. et al. (Jan. 2018). "Dynamic Reorganization of Open Chromatin Underlies Diverse Transcriptomes during Spermatogenesis". In: *Nucleic Acids Research* 46.2, pp. 593–608. DOI: 10.1093/nar/gkx1052 (cit. on p. 65).

Li, Y., P. Lemaire, and R. R. Behringer (Aug. 1, 1997). "Esx1,a Novel X Chromosome-Linked Homeobox Gene Expressed in Mouse Extraembryonic Tissues and Male Germ Cells". In: *Developmental Biology* 188.1, pp. 85–95. DOI: 10.1006/dbio.1997.8640 (cit. on p. 65).

Branford, W. W. et al. (July 1, 1997). "Spx1, a Novel X-Linked Homeobox Gene Expressed during Spermatogenesis". In: *Mechanisms of Development* 65.1, pp. 87–98. DOI: 10.1016/S0925-4773(97)00058-0 (cit. on p. 65).

Ding, X. et al. (June 2007). "SUN1 Is Required for Telomere Attachment to Nuclear Envelope and Gametogenesis in Mice". In: *Developmental cell* 12.6, pp. 863–872. DOI: 10.1016/j.devcel.2007.03.018. pmid: 17543860 (cit. on pp. 65, 117).

Tu, Z. et al. (Jan. 2017). "Speedy A-Cdk2 Binding Mediates Initial Telomere-Nuclear Envelope Attachment during Meiotic Prophase I Independent of Cdk2 Activation". In: *Proceedings of the National Academy of Sciences of the United States of America* 114.3, pp. 592–597. DOI: 10.1073/pnas.1618465114. pmid: 28031483 (cit. on pp. 65, 117).

Wang, Y. et al. (Feb. 2019). "The Meiotic TERB1-TERB2-MAJIN Complex Tethers Telomeres to the Nuclear Envelope". In: *Nature communications* 10.1, p. 564. DOI: 10.1038/s41467-019-08437-1. pmid: 30718482 (cit. on p. 65).

Lukaszewicz, A. et al. (July 2018). "Control of Meiotic Double-Strand-Break Formation by ATM: Local and Global Views". In: *Cell cycle* 17.10, pp. 1155–1172. DOI: 10.1080/15384101.2018.1464847. pmid: 29963942 (cit. on p. 65).

Reinholdt, L. G. and J. C. Schimenti (July 2005). "Mei1 Is Epistatic to Dmc1 during Mouse Meiosis". In: *Chromosoma* 114.2, pp. 127–134. DOI: 10.1007/s00412-005-0346-4. pmid: 15928951 (cit. on p. 65).

Brown, M. S. and D. K. Bishop (Dec. 2014). "DNA Strand Exchange and RecA Homologs in Meiosis". In: *Cold Spring Harbor perspectives in biology* 7.1, a016659. DOI: 10.1101/cshperspect.a016659. pmid: 25475089 (cit. on p. 65).

Dai, J. et al. (2017). "Meiotic Knockdown and Complementation Reveals Essential Role of RAD51 in Mouse Spermatogenesis". In: 18. DOI: 10.1016/j.celrep.2017.01.024 (cit. on p. 65).

Kovalenko, O. V., C. Wiese, and D. Schild (Sept. 2006). "RAD51AP2, a Novel Vertebrate- and Meiotic-Specific Protein, Shares a Conserved RAD51-Interacting C-Terminal Domain with RAD51AP1/PIR51". In: *Nucleic acids research* 34.18, pp. 5081–5092. DOI: 10.1093/nar/gkl665. pmid: 16990250 (cit. on p. 65).

Lee, K. Y. et al. (2015). "MCM8-9 Complex Promotes Resection of Double-Strand Break Ends by MRE11-RAD50-NBS1 Complex". In: 6. DOI: 10.1038/ncomms8744 (cit. on p. 65).

Martinez, J. S. et al. (Mar. 2016). "BRCA2 Regulates DMC1-Mediated Recombination through the BRC Repeats". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.13, pp. 3515–3520. DOI: 10.1073/pnas.1601691113. pmid: 26976601 (cit. on p. 65).

Pacheco, S. et al. (July 2018). "ATR Is Required to Complete Meiotic Recombination in Mice". In: *Nature communications* 9.1, p. 2622. DOI: 10.1038/s41467-018-04851-z. pmid: 29977027 (cit. on p. 65).

Ribeiro, J. et al. (June 2018). "MEIOB and SPATA22 Resemble RPA Subunits and Interact with the RPA Complex to Promote Meiotic Recombination". In: DOI: 10.1101/358242 (cit. on p. 65).

Widger, A. et al. (July 2018). "ATR Is a Multifunctional Regulator of Male Mouse Meiosis". In: *Nature communications* 9.1, p. 2621. DOI: 10.1038/s41467-018-04850-0. pmid: 29976923 (cit. on p. 65).

Yang, F. et al. (Feb. 2008). "Mouse TEX15 Is Essential for DNA Double-Strand Break Repair and Chromosomal Synapsis during Male Meiosis". In: *The Journal of cell biology* 180.4, pp. 673–679. DOI: 10.1083/jcb.200709057. pmid: 18283110 (cit. on p. 65).

Zhang, J. et al. (Feb. 2019). "A Meiosis-Specific BRCA2 Binding Protein Recruits Recombinases to DNA Double-Strand Breaks to Ensure Homologous Recombination". In: *Nature communications* 10.1, p. 722. DOI: 10.1038/s41467-019-08676-2. pmid: 30760716 (cit. on pp. 65, 66, 116).

Adelman, C. A. and J. H. J. Petrini (Mar. 2008). "ZIP4H (TEX11) Deficiency in the Mouse Impairs Meiotic Double Strand Break Repair and the Regulation of Crossing Over". In: *PLoS genetics* 4.3, e1000042. DOI: 10.1371/journal.pgen.1000042. pmid: 18369460 (cit. on p. 66).

Guiraldelli, M. F. et al. (2018). "SHOC1 Is a ERCC4-(HhH)2-like Protein, Integral to the Formation of Crossover Recombination Intermediates during Mammalian Meiosis". In: 14. DOI: 10.1371/journal.pgen.1007381 (cit. on p. 66).

Guiraldelli, M. F. et al. (Mar. 2013). "Mouse HFM1/Mer3 Is Required for Crossover Formation and Complete Synapsis of Homologous Chromosomes during Meiosis". In: *PLoS genetics* 9.3, e1003383. DOI: 10.1371/journal.pgen.1003383. pmid: 23555294 (cit. on p. 66).

Rakshambikai, R., N. Srinivasan, and K. T. Nishant (Nov. 2013). "Structural Insights into Saccharomyces Cerevisiae Msh4-Msh5 Complex Function Using Homology Modeling". In: *PloS one* 8.11, e78753. DOI: 10.1371/journal.pone.0078753. pmid: 24244354 (cit. on p. 66).

Sun, X. et al. (2016). "FancJ (Brip1) Loss-of-Function Allele Results in Spermatogonial Cell Depletion during Embryogenesis and Altered Processing of Crossover Sites during Meiotic Prophase I in Mice". In: 125. DOI: 10.1007/s00412-015-0549-2 (cit. on p. 66).

Gómez-H, L. et al. (Oct. 2016). "C14ORF39/SIX6OS1 Is a Constituent of the Synaptonemal Complex and Is Essential for Mouse Fertility". In: *Nature communications* 7, p. 13298. DOI: 10.1038/ncomms13298. pmid: 27796301 (cit. on p. 66).

Syrjänen, J. L., L. Pellegrini, and O. R. Davies (June 2014). "A Molecular Model for the Role of SYCP3 in Meiotic Chromosome Organisation". In: *eLife* 3. DOI: 10.7554/eLife.02963. pmid: 24950965 (cit. on p. 66).

Abby, E. et al. (Jan. 2016). "Implementation of Meiosis Prophase I Programme Requires a Conserved Retinoid-Independent Stabilizer of Meiotic Transcripts". In: *Nature communications* 7, p. 10324. DOI: 10.1038/ncomms10324. pmid: 26742488 (cit. on p. 66).

Soh, Y. Q. S. et al. (Apr. 2017). "Meioc Maintains an Extended Meiotic Prophase I in Mice". In: *PLoS genetics* 13.4, e1006704. DOI: 10.1371/journal.pgen.1006704. pmid: 28380054 (cit. on p. 66).

Boekhout, M. et al. (2018). "REC114 Partner ANKRD31 Controls Number, Timing and Location of Meiotic DNA Breaks". In: DOI: 10.1101/425322 (cit. on p. 66).

Spruce, C. et al. (2019). "HELLS and PRDM9 Form a Pioneer Complex to Open Chromatin at Meiotic Recombination Hotspots". In: *bioRxiv* (cit. on pp. 67, 117).

Kleene, K. C. (Mar. 2016). "Position-Dependent Interactions of Y-Box Protein 2 (YBX2) with mRNA Enable mRNA Storage in Round Spermatids by Repressing mRNA Translation and Blocking Translation-Dependent mRNA Decay". In: *Molecular Reproduction and Development* 83.3, pp. 190–207. DOI: 10.1002/mrd.22616 (cit. on p. 67).

Kwon, Y. K., M. T. Murray, and N. B. Hecht (July 1, 1993). "Proteins Homologous to the Xenopus Germ Cell-Specific RNA-Binding Proteins P54/P56 Are Temporally Expressed in Mouse Male Germ Cells". In: *Developmental Biology* 158.1, pp. 90–100. DOI: 10.1006/dbio.1993.1170 (cit. on p. 67).

Oko, R. et al. (1996). "Germ Cell-Specific DNA and RNA Binding Proteins P48/52 Are Expressed at Specific Stages of Male Germ Cell Development and Are Present in the Chromatoid Body". In: *Molecular Reproduction and Development* 44.1, pp. 1–13. DOI: 10.1002/(SICI)1098-2795(199605)44:1<1::AID-MRD1>3.0.CO;2-S (cit. on p. 67).

Gu, W. et al. (Nov. 1, 1998). "Mammalian Male and Female Germ Cells Express a Germ Cell-Specific Y-Box Protein, MSY2". In: *Biology of Reproduction* 59.5, pp. 1266–1274. DOI: 10.1095/biolreprod59.5.1266 (cit. on p. 67).

Yang, J. et al. (Apr. 19, 2005). "Absence of the DNA-/RNA-Binding Protein MSY2 Results in Male and Female Infertility". In: *Proceedings of the National Academy of Sciences* 102.16, pp. 5755–5760. DOI: 10.1073/pnas.0408718102. pmid: 15824319 (cit. on p. 67).

Tafuri, S. R., M. Familari, and A. P. Wolffe (May 6, 1993). "A Mouse Y Box Protein, MSY1, Is Associated with Paternal mRNA in Spermatocytes." In: *Journal of Biological Chemistry* 268.16, pp. 12213–12220. pmid: 8505341 (cit. on p. 67).

Davies, H. G. et al. (May 1, 2000). "A Sequence-Specific RNA Binding Complex Expressed in Murine Germ Cells Contains MSY2 and MSY4". In: *Developmental Biology* 221.1, pp. 87–100. DOI: 10.1006/dbio.2000.9658 (cit. on p. 67).

Giorgini, F., H. G. Davies, and R. E. Braun (Aug. 1, 2002). "Translational Repression by MSY4 Inhibits Spermatid Differentiation in Mice". In: *Development* 129.15, pp. 3669–3679. pmid: 12117816 (cit. on p. 67).

Ozturk, S. and F. Uysal (Aug. 2018). "Potential Roles of the Poly(A)-Binding Proteins in Translational Regulation during Spermatogenesis". In: *The Journal of Reproduction and Development* 64.4, pp. 289–296. DOI: 10.1262/jrd.2018-026. pmid: 29780056 (cit. on p. 67).

C. Kleene, K. et al. (Dec. 1, 1994). "Developmental Expression of Poly(A) Binding Protein mRNAs during Spermatogenesis in the Mouse". In: *Molecular Reproduction and Development* 39.4, pp. 355–364. DOI: 10.1002/mrd.1080390403 (cit. on p. 67).

Gu, W. et al. (1995). "Poly (A) Binding Protein Is Bound to Both Stored and Polysomal mRNAs in the Mammalian Testis". In: *Molecular Reproduction and Development* 40.3, pp. 273–285. DOI: 10.1002/mrd.1080400303 (cit. on p. 67).

Kleene, K. C. et al. (Sept. 1, 1998). "The Mouse Gene Encoding the Testis-Specific Isoform of Poly(A) Binding Protein (Pabp2) Is an Expressed Retroposon: Intimations That Gene Expression in Spermatogenic Cells Facilitates the Creation of New Genes". In: *Journal of Molecular Evolution* 47.3, pp. 275–281. DOI: 10.1007/PL00006385 (cit. on p. 67).

Lee, Y. J., J. K. Ahn, and J. H. Chung (2000). "Expression of Murine Poly(A) Binding Protein II Gene in Testis". In: *IUBMB Life* 49.1, pp. 57–61. DOI: 10.1080/713803589 (cit. on p. 67).

Kimura, M., K. Ishida, et al. (Mar. 2009). "Characterization of Two Cytoplasmic Poly(A)-Binding Proteins, PABPC1 and PABPC2, in Mouse Spermatogenic Cells1". In: *Biology of Reproduction* 80.3, pp. 545–554. DOI: 10.1095/biolreprod.108.072553 (cit. on p. 67).

Goldberg, E. (Feb. 15, 1963). "Lactic and Malic Dehydrogenases in Human Spermatozoa". In: *Science* 139.3555, pp. 602–603. DOI: 10.1126/science.139.3555.602. pmid: 17788301 (cit. on p. 67).

Blanco, A. and W. H. Zinkham (Feb. 15, 1963). "Lactate Dehydrogenases in Human Testes". In: *Science* 139.3555, pp. 601–602. DOI: 10.1126/science.139.3555.601. pmid: 17788300 (cit. on p. 67).

Goldberg, E. et al. (2010). "LDHC THE ULTIMATE TESTIS SPECIFIC GENE". In: *Journal of andrology* 31.1, pp. 86–94. DOI: 10.2164/jandrol.109.008367. pmid: 19875487 (cit. on p. 67).

Wang, H. et al. (June 1, 2005). "Cloning and Characterization of a Novel Intronless Lactate Dehydrogenase Gene in Human Testis". In: *International Journal of Molecular Medicine* 15.6, pp. 949–953. DOI: 10.3892/ijmm.15.6.949 (cit. on p. 67).

Fundele, R. et al. (Dec. 1, 1987). "Developmental Activation of Phosphoglycerate Mutase-2 in the Testis of the Mouse". In: *Developmental Biology* 124.2, pp. 562–566. DOI: 10.1016/0012-1606(87)90510-0 (cit. on pp. 67, 82).

Gou, L.-T. T. et al. (June 2014). "Pachytene piRNAs Instruct Massive mRNA Elimination during Late Spermiogenesis." In: *Cell research* 24.6, pp. 680–700. DOI: 10.1038/cr.2014.41. pmid: 24787618 (cit. on p. 68).

Iwasaki, Y. W., M. C. Siomi, and H. Siomi (2015). "PIWI-Interacting RNA: Its Biogenesis and Functions". In: *Annual Review of Biochemistry* 84.1, pp. 405–433. DOI: 10.1146/annurev-biochem-060614-034258. pmid: 25747396 (cit. on p. 68).

Dun, M. D., R. J. Aitken, and B. Nixon (July 1, 2012). "The Role of Molecular Chaperones in Spermatogenesis and the Post-Testicular Maturation of Mammalian Spermatozoa". In: *Human Reproduction Update* 18.4, pp. 420–435. DOI: 10.1093/humupd/dms009 (cit. on p. 68).

Turner, J. M. A. (Apr. 2007). "Meiotic Sex Chromosome Inactivation". In: *Development* 134.10, pp. 1823–1831. DOI: 10.1242/dev.000018. pmid: 17329371 (cit. on p. 68).

Turner, J. M. A. (2015). "Meiotic Silencing in Mammals". In: *Annual review of genetics* 49.1, pp. 395–412. DOI: 10.1146/annurev-genet-112414-055145 (cit. on p. 68).

Greenbaum, M. P. et al. (Aug. 2011). "Germ Cell Intercellular Bridges". In: *Cold Spring Harbor perspectives in biology* 3.8, a005850. DOI: 10.1101/cshperspect.a005850. pmid: 21669984 (cit. on p. 70).

Ren, H. P. and L. D. Russell (1991). "Clonal Development of Interconnected Germ Cells in the Rat and Its Relationship to the Segmental and Subsegmental Organization of Spermatogenesis". In: *American journal of anatomy* 192.2, pp. 121–128. DOI: 10.1002/aja.1001920203 (cit. on p. 70).

Braun, R. E. et al. (1989). "Genetically Haploid Spermatids Are Phenotypically Diploid". In: *Nature* 337.6205, pp. 373–376. DOI: 10.1038/337373a0 (cit. on p. 70).

Herrmann, B. G. et al. (Nov. 1999). "A Protein Kinase Encoded by the t Complex Responder Gene Causes Non-Mendelian Inheritance". In: *Nature* 402.6758, pp. 141–146. DOI: 10.1038/45970 (cit. on p. 71).

Véron, N. et al. (2009). "Retention of Gene Products in Syncytial Spermatids Promotes Non-Mendelian Inheritance as Revealed by the t Complex Responder". In: *Genes & Development* 23.23, pp. 2705–2710. DOI: 10.1101/gad.553009 (cit. on p. 71).

Martin-DeLeon, P. A. et al. (2005). "Spam1-Associated Transmission Ratio Distortion in Mice: Elucidating the Mechanism". In: *Reproductive Biology and Endocrinology* 3. DOI: 10.1186/1477-7827-3-32 (cit. on p. 71).

Bhutani, K. et al. (Nov. 18, 2019). "Widespread Haploid-Biased Gene Expression in Mammalian Spermatogenesis Associated with Frequent Selective Sweeps and Evolutionary Conflict". In: *bioRxiv*, p. 846253. DOI: 10.1101/846253 (cit. on p. 71).

Royo, H. et al. (Dec. 7, 2010). "Evidence That Meiotic Sex Chromosome Inactivation Is Essential for Male Fertility". In: *Current Biology* 20.23, pp. 2117–2123. DOI: 10.1016/j.cub.2010.11.010 (cit. on p. 71).

Ortega, M. M. (2016). "Surveillance Mechanisms in Mammalian Meiosis". PhD Thesis. Universitat Autònoma de Barcelona (cit. on p. 71).

Wojtasz, L. et al. (Oct. 2009). "Mouse HORMAD1 and HORMAD2, Two Conserved Meiotic Chromosomal Proteins, Are Depleted from Synapsed Chromosome Axes with the Help of TRIP13 AAA-ATPase". In: *PLoS genetics* 5.10, e1000702. DOI: 10.1371/journal.pgen.1000702. pmid: 19851446 (cit. on p. 71).

Dubiel, W. et al. (Feb. 2018). "Cullin 3-Based Ubiquitin Ligases as Master Regulators of Mammalian Cell Differentiation". In: *Trends in biochemical sciences* 43.2, pp. 95–107. DOI: 10.1016/j.tibs.2017.11.010. pmid: 29249570 (cit. on p. 74).

Sweeney, C. et al. (Jan. 1, 1996). "A Distinct Cyclin A Is Expressed in Germ Cells in the Mouse". In: *Development* 122.1, pp. 53–64. pmid: 8565853 (cit. on p. 75).

Liu, D. et al. (Dec. 1998). "Cyclin A1 Is Required for Meiosis in the Male Mouse". In: *Nature Genetics* 20.4, pp. 377–380. DOI: 10.1038/3855 (cit. on p. 75).

Joukov, V. and A. D. Nicolo (Aug. 14, 2018). "Aurora-PLK1 Cascades as Key Signaling Modules in the Regulation of Mitosis". In: *Science Signaling* 11.543. DOI: 10.1126/scisignal.aar4195. pmid: 30108183 (cit. on p. 75).

Poon, R. Y. C. and T. Hunter (Oct. 6, 1995). "Dephosphorylation of Cdk2 Thr160 by the Cyclin-Dependent Kinase-Interacting Phosphatase KAP in the Absence of Cyclin". In: *Science* 270.5233, pp. 90–93. DOI: 10.1126/science.270.5233.90. pmid: 7569954 (cit. on p. 75).

Saigusa, K. et al. (Feb. 2007). "RGC32, a Novel P53-Inducible Gene, Is Located on Centrosomes during Mitosis and Results in G2/M Arrest". In: *Oncogene* 26.8, pp. 1110–1121. DOI: 10.1038/sj.onc.1210148 (cit. on p. 75).

Holt, J. E. et al. (Mar. 15, 2014). "The APC/C Activator FZR1 Is Essential for Meiotic Prophase I in Mice". In: *Development* 141.6, pp. 1354–1365. DOI: 10.1242/dev.104828. pmid: 24553289 (cit. on p. 75).

Habu, T. et al. (Dec. 2, 2002). "Identification of a MAD2-Binding Protein, CMT2, and Its Role in Mitosis". In: *The EMBO Journal* 21.23, pp. 6419–6428. DOI: 10.1093/emboj/cdf659. pmid: 12456649 (cit. on p. 75).

Church, D. M. et al. (May 2009). "Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse". In: *PLoS biology* 7.5, e1000112. DOI: 10.1371/journal.pbio.1000112. pmid: 19468303 (cit. on p. 76).

Moretti, C. et al. (Oct. 2016). "Expression and Epigenomic Landscape of the Sex Chromosomes in Mouse Post-Meiotic Male Germ Cells". In: *Epigenetics & chromatin* 9, p. 47. DOI: 10.1186/s13072-016-0099-8. pmid: 27795737 (cit. on p. 76).

Ito, C. and K. Toshimori (Mar. 2016). "Acrosome Markers of Human Sperm". In: *Anatomical science international* 91.2, pp. 128–142. DOI: 10.1007/s12565-015-0323-9. pmid: 26748928 (cit. on p. 76).

Hao, Z. et al. (Mar. 1, 2002). "SAMP32, a Testis-Specific, Isoantigenic Sperm Acrosomal Membrane-Associated Protein". In: *Biology of Reproduction* 66.3, pp. 735–744. DOI: 10.1095/biolreprod66.3.735 (cit. on p. 76).

Fujihara, Y. et al. (Oct. 1, 2012). "SPACA1-Deficient Male Mice Are Infertile with Abnormally Shaped Sperm Heads Reminiscent of Globozoospermia". In: *Development* 139.19, pp. 3583–3589. DOI: 10.1242/dev.081778. pmid: 22949614 (cit. on p. 76).

Wei, J. et al. (Nov. 2013). "Characterisation of Lyzls in Mice and Antibacterial Properties of Human LYZL6". In: *Asian Journal of Andrology* 15.6, pp. 824–830. DOI: 10.1038/aja.2013.93 (cit. on p. 76).

Mandal, A. et al. (May 1, 2003). "SLLP1, A Unique, Intra-Acrosomal, Non-Bacteriolytic, c Lysozyme-Like Protein of Human Spermatozoa". In: *Biology of Reproduction* 68.5, pp. 1525–1537. DOI: 10.1095/biolreprod.102.010108 (cit. on p. 76).

Herrero, M. B. et al. (Aug. 1, 2005). "Mouse SLLP1, a Sperm Lysozyme-like Protein Involved in Sperm–Egg Binding and Fertilization". In: *Developmental Biology* 284.1, pp. 126–142. DOI: 10.1016/j.ydbio.2005.05.008 (cit. on pp. 76, 77).

Guyonnet, B. et al. (Sept. 2012). "Isolation and Proteomic Characterization of the Mouse Sperm Acrosomal Matrix". In: *Molecular & Cellular Proteomics : MCP* 11.9, p. 758. DOI: 10.1074/mcp.M112.020339. pmid: 22707618 (cit. on p. 77).

Sun, R. et al. (May 1, 2011). "Lyzl4, a Novel Mouse Sperm-Related Protein, Is Involved in Fertilization". In: *Acta Biochimica et Biophysica Sinica* 43.5, pp. 346–353. DOI: 10.1093/abbs/gmr017 (cit. on p. 77).

Zhang, K. et al. (Nov. 1, 2005). "Molecular Cloning and Characterization of Three Novel Lysozyme-Like Genes, Predominantly Expressed in the Male Reproductive System of Humans, Belonging to the C-Type Lysozyme/Alpha-Lactalbumin Family". In: *Biology of Reproduction* 73.5, pp. 1064–1071. DOI: 10.1095/biolreprod.105.041889 (cit. on p. 77).

Korfanty, J. et al. (June 1, 2012). "Identification of a New Mouse Sperm Acrosome-Associated Protein". In: *Reproduction* 143.6, pp. 749–757. DOI: 10.1530/REP-11-0270 (cit. on p. 77).

Nguyen, E. B., A. D. Westmuckett, and K. L. Moore (Jan. 2014). "SPACA7 Is a Novel Male Germ Cell-Specific Protein Localized to the Sperm Acrosome That Is Involved in Fertilization in Mice". In: *Biology of Reproduction* 90.1, p. 16. DOI: 10.1095/biolreprod.113.111831. pmid: 24307706 (cit. on p. 77).

Shetty, J. et al. (Aug. 15, 2003). "SAMP14, a Novel, Acrosomal Membrane-Associated, Glycosylphosphatidylinositol-Anchored Member of the Ly-6/Urokinase-Type Plasminogen Activator Receptor Superfamily with a Role in Sperm-Egg Interaction". In: *Journal of Biological Chemistry* 278.33, pp. 30506–30515. DOI: 10.1074/jbc.M301713200. pmid: 12788941 (cit. on p. 77).

Anway, M. D. et al. (2003). "Expression of Testicular Germ Cell Genes Identified by Differential Display Analysis". In: *Journal of Andrology* 24.2, pp. 173–184. DOI: 10.1002/j.1939-4640.2003.tb02660.x (cit. on p. 77).

Li, Y. et al. (2015). "FAM170B, a Novel Acrosomal Protein Involved in Fertilization in Mice". In: *Molecular Reproduction and Development* 82.10, pp. 787–796. DOI: 10.1002/mrd.22523 (cit. on p. 77).

Luk, J. M. et al. (2006). "Acrosome-Specific Gene AEP1: Identification, Characterization and Roles in Spermatogenesis". In: *Journal of Cellular Physiology* 209.3, pp. 755–766. DOI: 10.1002/jcp.20746 (cit. on p. 77).

Chen, J. et al. (Sept. 1, 2016). "Deficiency of SPATA46, a Novel Nuclear Membrane Protein, Causes Subfertility in Male Mice". In: *Biology of Reproduction* 95.3. DOI: 10.1095/biolreprod.116.140996 (cit. on p. 77).

Myles, D. G., P. Primakoff, and A. R. Bellvé (Feb. 1981). "Surface Domains of the Guinea Pig Sperm Defined with Monoclonal Antibodies". In: *Cell* 23.2, pp. 433–439. DOI: 10.1016/0092-8674(81)90138-0. pmid: 7008952 (cit. on p. 78).

Primakoff, P., H. Hyatt, and D. G. Myles (Dec. 1, 1985). "A Role for the Migrating Sperm Surface Antigen PH-20 in Guinea Pig Sperm Binding to the Egg Zona Pellucida." In: *The Journal of Cell Biology* 101.6, pp. 2239–2244. DOI: 10.1083/jcb.101.6.2239. pmid: 4066757 (cit. on p. 78).

Lathrop, W. F. et al. (Dec. 1, 1990). "cDNA Cloning Reveals the Molecular Structure of a Sperm Surface Protein, PH-20, Involved in Sperm-Egg Adhesion and the Wide Distribution of Its Gene among Mammals." In: *The Journal of Cell Biology* 111.6, pp. 2939–2949. DOI: 10.1083/jcb.111.6.2939. pmid: 2269661 (cit. on p. 78).

Baba, D. et al. (Aug. 16, 2002). "Mouse Sperm Lacking Cell Surface Hyaluronidase PH-20 Can Pass through the Layer of Cumulus Cells and Fertilize the Egg". In: *Journal of Biological Chemistry* 277.33, pp. 30310–30314. DOI: 10.1074/jbc.M204596200. pmid: 12065596 (cit. on p. 78).

Kim, E. et al. (Dec. 13, 2005). "Identification of a Hyaluronidase, Hyal5, Involved in Penetration of Mouse Sperm through Cumulus Mass". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.50, pp. 18028–18033. DOI: 10.1073/pnas.0506825102. pmid: 16330764 (cit. on p. 78).

Kimura, M., E. Kim, et al. (Nov. 1, 2009). "Functional Roles of Mouse Sperm Hyaluronidases, HYAL5 and SPAM1, in Fertilization". In: *Biology of Reproduction* 81.5, pp. 939–947. DOI: 10.1095/biolreprod.109.078816 (cit. on p. 78).

Li, Y.-C. et al. (2006). "Afaf, a Novel Vesicle Membrane Protein, Is Related to Acrosome Formation in Murine Testis". In: *FEBS Letters* 580.17, pp. 4266–4273. DOI: 10.1016/j.febslet.2006.06.010 (cit. on p. 78).

Hao, J. et al. (Feb. 21, 2014). "Equatorin Is Not Essential for Acrosome Biogenesis but Is Required for the Acrosome Reaction". In: *Biochemical and Biophysical Research Communications* 444.4, pp. 537–542. DOI: 10.1016/j.bbrc.2014.01.080 (cit. on p. 78).

Inoue, N. et al. (Mar. 2005). "The Immunoglobulin Superfamily Protein Izumo Is Required for Sperm to Fuse with Eggs". In: *Nature* 434.7030, pp. 234–238. DOI: 10.1038/nature03362 (cit. on p. 78).

Bianchi, E. et al. (Apr. 2014). "Juno Is the Egg Izumo Receptor and Is Essential for Mammalian Fertilization". In: *Nature* 508.7497, pp. 483–487. DOI: 10.1038/nature13203 (cit. on p. 78).

Nishimura, H. et al. (Apr. 1, 2011). "Characterization of Mouse Sperm TMEM190, a Small Transmembrane Protein with the Trefoil Domain: Evidence for Co-Localization with IZUMO1 and Complex Formation with Other Sperm Proteins". In: *Reproduction* 141.4, pp. 437–451. DOI: 10.1530/REP-10-0391 (cit. on p. 78).

Gibbs, G. M. et al. (May 1, 2010). "Glioma Pathogenesis-Related 1-Like 1 Is Testis Enriched, Dynamically Modified, and Redistributed during Male Germ Cell Maturation and Has a Potential Role in Sperm-Oocyte Binding". In: *Endocrinology* 151.5, pp. 2331–2342. DOI: 10.1210/en.2009-1255 (cit. on p. 78).

Gaikwad, A. S. et al. (Oct. 31, 2019). "GLIPR1L1 Is an IZUMO-Binding Protein Required for Optimal Fertilization in the Mouse". In: *BMC Biology* 17. DOI: 10.1186/s12915-019-0701-1. pmid: 31672133 (cit. on p. 78).

Bleil, J. D. and P. M. Wassarman (July 1, 1990). "Identification of a ZP3-Binding Protein on Acrosome-Intact Mouse Sperm by Photoaffinity Crosslinking." In: *Proceedings of the National Academy of Sciences* 87.14, pp. 5563–5567. DOI: 10.1073/pnas.87.14.5563. pmid: 2371290 (cit. on p. 78).

Kim, K.-S., M. C. Cha, and G. L. Gerton (Jan. 1, 2001). "Mouse Sperm Protein Sp56 Is a Component of the Acrosomal Matrix". In: *Biology of Reproduction* 64.1, pp. 36–43. DOI: 10.1095/biolreprod64.1.36 (cit. on p. 78).

Muro, Y. et al. (Jan. 1, 2012). "Function of the Acrosomal Matrix: Zona Pellucida 3 Receptor (ZP3R/Sp56) Is Not Essential for Mouse Fertilization". In: *Biology of Reproduction* 86.1. DOI: 10.1095/biolreprod.111.095877 (cit. on p. 78).

Lin, Y.-N. et al. (Oct. 2007). "Loss of Zona Pellucida Binding Proteins in the Acrosomal Matrix Disrupts Acrosome Biogenesis and Sperm Morphogenesis". In: *Molecular and Cellular Biology* 27.19, pp. 6794–6805. DOI: 10.1128/MCB.01029-07. pmid: 17664285 (cit. on p. 78).

Ren, D. et al. (Oct. 2001). "A Sperm Ion Channel Required for Sperm Motility and Male Fertility". In: *Nature* 413.6856, pp. 603–609. DOI: 10.1038/35098027 (cit. on p. 78).

Lobley, A. et al. (Aug. 1, 2003). "Identification of Human and Mouse CatSper3 and CatSper4 Genes: Characterisation of a Common Interaction Domain and Evidence for Expression in Testis". In: *Reproductive Biology and Endocrinology* 1.1, p. 53. DOI: 10.1186/1477-7827-1-53 (cit. on p. 78).

Qi, H. et al. (Jan. 23, 2007). "All Four CatSper Ion Channel Proteins Are Required for Male Fertility and Sperm Cell Hyperactivated Motility". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.4, pp. 1219–1223. DOI: 10.1073/pnas.0610286104. pmid: 17227845 (cit. on p. 78).

Jin, J. et al. (July 1, 2007). "Catsper3 and Catsper4 Are Essential for Sperm Hyperactivated Motility and Male Fertility in the Mouse". In: *Biology of Reproduction* 77.1, pp. 37–44. DOI: 10.1095/biolreprod.107.060186 (cit. on p. 78).

Chung, J.-J. et al. (Feb. 22, 2017). "CatSperζ Regulates the Structural Continuity of Sperm Ca2+ Signaling Domains and Is Required for Normal Fertility". In: *eLife* 6. Ed. by K. J. Swartz, e23082. DOI: 10.7554/eLife.23082 (cit. on p. 78).

Choi, I. et al. (Apr. 1, 2004). "Characterization and Comparative Genomic Analysis of Intronless Adams with Testicular Gene Expression". In: *Genomics* 83.4, pp. 636–646. DOI: 10.1016/j.ygeno.2003.10.001 (cit. on p. 79).

Zhu, G.-Z. et al. (2009). "Testase 1 (ADAM 24) a sperm surface metalloprotease is required for normal fertility in mice". In: *Molecular Reproduction and Development* 76.11, pp. 1106–1114. DOI: 10.1002/mrd.21076 (cit. on p. 79).

Amos, L. A. (2008). "The Tektin Family of Microtubule-Stabilizing Proteins". In: *Genome Biology* 9.7, p. 229. DOI: 10.1186/gb-2008-9-7-229. pmid: 18671835 (cit. on p. 79).

Fallahi, M. et al. (Dec. 13, 2010). "A Global Expression Switch Marks Pachytene Initiation during Mouse Male Meiosis". In: *Genes* 1.3, pp. 469–483. DOI: 10.3390/genes1030469. pmid: 24710097 (cit. on p. 79).

Li, Y. et al. (Jan. 2011). "Expression and Localization of Five Members of the Testis-Specific Serine Kinase (Tssk) Family in Mouse and Human Sperm and Testis". In: *Molecular Human Reproduction* 17.1, pp. 42–56. DOI: 10.1093/molehr/gaq071. pmid: 20729278 (cit. on p. 79).

Shen, C. et al. (May 1, 2013). "Prss37 Is Required for Male Fertility in the Mouse". In: *Biology of Reproduction* 88.5. DOI: 10.1095/biolreprod.112.107086 (cit. on p. 80).

Petit, F. G. et al. (Mar. 1, 2015). "Combining RNA and Protein Profiling Data with Network Interactions Identifies Genes Associated with Spermatogenesis in Mouse and Human". In: *Biology of Reproduction* 92.3. DOI: 10.1095/biolreprod.114.126250 (cit. on p. 80).

Van Koningsbruggen, S. et al. (Feb. 1, 2007). "FRG1P-Mediated Aggregation of Proteins Involved in Pre-mRNA Processing". In: *Chromosoma* 116.1, pp. 53–64. DOI: 10.1007/s00412-006-0083-3 (cit. on p. 80).

Cummerson, J. A. et al. (July 1, 2006). "The Complement Regulatory Proteins CD55 (Decay Accelerating Factor) and CD59 Are Expressed on the Inner Acrosomal Membrane of Human Spermatozoa as Well as CD46 (Membrane Cofactor Protein)". In: *Immunology* 118.3, pp. 333–342. DOI: 10.1111/j.1365-2567.2006.02374.x (cit. on p. 80).

Inoue, N. et al. (Apr. 1, 2003). "Disruption of Mouse CD46 Causes an Accelerated Spontaneous Acrosome Reaction in Sperm". In: *Molecular and Cellular Biology* 23.7, pp. 2614–2622. DOI: 10.1128/MCB.23.7.2614-2622.2003. pmid: 12640142 (cit. on p. 80).

Kaji, K. et al. (Mar. 2000). "The Gamete Fusion Process Is Defective in Eggs of Cd9-Deficient Mice". In: *Nature Genetics* 24.3, pp. 279–282. DOI: 10.1038/73502 (cit. on p. 81).

Naour, F. L. et al. (Jan. 14, 2000). "Severely Reduced Female Fertility in CD9-Deficient Mice". In: *Science* 287.5451, pp. 319–321. DOI: 10.1126/science.287.5451.319. pmid: 10634790 (cit. on p. 81).

Miyado, K. et al. (Jan. 14, 2000). "Requirement of CD9 on the Egg Plasma Membrane for Fertilization". In: *Science* 287.5451, pp. 321–324. DOI: 10.1126/science.287.5451.321. pmid: 10634791 (cit. on p. 81).

Knobeloch, K.-P. et al. (Aug. 2000). "Targeted Inactivation of the Tetraspanin CD37 Impairs T-Cell-Dependent B-Cell Response under Suboptimal Costimulatory Conditions". In: *Molecular and Cellular Biology* 20.15, pp. 5363–5369. pmid: 10891477 (cit. on p. 81).

Kwon, J. T. et al. (July 25, 2017). "Expression of Uncharacterized Male Germ Cell-Specific Genes and Discovery of Novel Sperm-Tail Proteins in Mice". In: *PLOS ONE* 12.7, e0182038. DOI: 10.1371/journal.pone.0182038 (cit. on p. 81).

Sassone-Corsi, P. (June 2002). "Unique Chromatin Remodeling and Transcriptional Regulation in Spermatogenesis". In: *Science* 296.5576, pp. 2176–2178. DOI: 10.1126/science.1070963. pmid: 12077401 (cit. on p. 81).

Martianov, I. et al. (Feb. 22, 2005). "Polar Nuclear Localization of H1T2, a Histone H1 Variant, Required for Spermatid Elongation and DNA Condensation during Spermiogenesis". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.8, pp. 2808–2813. DOI: 10.1073/pnas.0406060102. pmid: 15710904 (cit. on pp. 81, 94).

Tanaka, H. et al. (Aug. 2005). "HANP1/H1T2, a Novel Histone H1-Like Protein Involved in Nuclear Formation and Sperm Fertility". In: *Molecular and Cellular Biology* 25.16, pp. 7107–7119. DOI: 10.1128/MCB.25.16.7107-7119.2005. pmid: 16055721 (cit. on p. 81).

Yan, W. et al. (Sept. 2, 2003). "HILS1 Is a Spermatid-Specific Linker Histone H1-like Protein Implicated in Chromatin Remodeling during Mammalian Spermiogenesis". In: *Proceedings of the National Academy of Sciences of the United States of America* 100.18, pp. 10546–10551. DOI: 10.1073/pnas.1837812100. pmid: 12920187 (cit. on p. 81).

Wu, Q. et al. (July 1, 2009). "Genetic Interaction of Hils1 with Tnp1 Is Essential for Normal Sperm Production and Male Fertility." In: *Biology of Reproduction* 81 (Suppl_1), pp. 462–462. DOI: 10.1093/biolreprod/81.s1.462 (cit. on p. 81).

Westhoff, D. and G. Kamp (Aug. 1, 1997). "Glyceraldehyde 3-Phosphate Dehydrogenase Is Bound to the Fibrous Sheath of Mammalian Spermatozoa". In: *Journal of Cell Science* 110.15, pp. 1821–1829. pmid: 9264469 (cit. on p. 82).

Krisfalusi, M. et al. (Aug. 1, 2006). "Multiple Glycolytic Enzymes Are Tightly Bound to the Fibrous Sheath of Mouse Spermatozoa". In: *Biology of Reproduction* 75.2, pp. 270–278. DOI: 10.1095/biolreprod.105.049684 (cit. on p. 82).

Mukai, C. and M. Okuno (Aug. 1, 2004). "Glycolysis Plays a Major Role for Adenosine Triphosphate Supplementation in Mouse Sperm Flagellar Movement". In: *Biology of Reproduction* 71.2, pp. 540–547. DOI: 10.1095/biolreprod.103.026054 (cit. on p. 82).

Mori, C. et al. (Aug. 1, 1993). "Unique Hexokinase Messenger Ribonucleic Acids Lacking the Porin-Binding Domain Are Developmentally Expressed in Mouse Spermatogenic Cells". In: *Biology of Reproduction* 49.2, pp. 191–203. DOI: 10.1095/biolreprod49.2.191 (cit. on p. 82).

Nakamura, N. et al. (Apr. 2008). "Spermatogenic Cell-Specific Type 1 Hexokinase Is the Predominant Hexokinase in Sperm". In: *Molecular reproduction and development* 75.4, pp. 632–640. DOI: 10.1002/mrd.20791. pmid: 17924400 (cit. on p. 82).

Buehr, M. and A. McLaren (Sept. 1, 1981). "An Electrophoretically Detectable Modification of Glucosephosphate Isomerase in Mouse Spermatozoa". In: *Reproduction* 63.1, pp. 169–173. DOI: 10.1530/jrf.0.0630169 (cit. on p. 82).

Vemuganti, S. A., F. P.-M. de Villena, and D. A. O'Brien (May 6, 2010). "Frequent and Recent Retrotransposition of Orthologous Genes Plays a Role in the Evolution of Sperm Glycolytic Enzymes". In: *BMC Genomics* 11.1, p. 285. DOI: 10.1186/1471-2164-11-285 (cit. on p. 82).

Vemuganti, S. A. et al. (Sept. 1, 2007). "Three Male Germline-Specific Aldolase A Isozymes Are Generated by Alternative Splicing and Retrotransposition". In: *Developmental Biology* 309.1, pp. 18–31. DOI: 10.1016/j.ydbio.2007.06.010 (cit. on p. 82).

Ijiri, T. W. et al. (2013). "Male Mice Express Spermatogenic Cell-Specific Triosephosphate Isomerase Isozymes". In: *Molecular Reproduction and Development* 80.10, pp. 862–870. DOI: 10.1002/mrd.22217 (cit. on p. 82).

McCarrey, J. R. and K. Thomas (Apr. 1987). "Human Testis-Specific PGK Gene Lacks Introns and Possesses Characteristics of a Processed Gene". In: *Nature* 326.6112, pp. 501–505. DOI: 10.1038/326501a0 (cit. on p. 82).

Welch, J. E. et al. (May 1, 1992). "Expression of a Glyceraldehyde 3-Phosphate Dehydrogenase Gene Specific to Mouse Spermatogenic Cells". In: *Biology of Reproduction* 46.5, pp. 869–878. DOI: 10.1095/biolreprod46.5.869 (cit. on p. 82).

Miki, K. et al. (Nov. 23, 2004). "Glyceraldehyde 3-Phosphate Dehydrogenase-S, a Sperm-Specific Glycolytic Enzyme, Is Required for Sperm Motility and Male Fertility". In: *Proceedings of the National Academy of Sciences* 101.47, pp. 16501–16506. DOI: 10.1073/pnas.0407708101. pmid: 15546993 (cit. on p. 82).

Danshina, P. V. et al. (Jan. 1, 2010). "Phosphoglycerate Kinase 2 (PGK2) Is Essential for Sperm Function and Male Fertility in Mice". In: *Biology of Reproduction* 82.1, pp. 136–145. DOI: 10.1095/biolreprod.109.079699 (cit. on p. 82).

Sakai, I., F. S. Sharief, and S. S. L. Li (Mar. 1, 1987). "Molecular Cloning and Nucleotide Sequence of the cDNA for Sperm-Specific Lactate Dehydrogenase-C from Mouse". In: *Biochemical Journal* 242.2, pp. 619–622. DOI: 10.1042/bj2420619 (cit. on p. 83).

Odet, F. et al. (July 1, 2008). "Expression of the Gene for Mouse Lactate Dehydrogenase C (Ldhc) Is Required for Male Fertility". In: *Biology of Reproduction* 79.1, pp. 26–34. DOI: 10.1095/biolreprod.108.068353 (cit. on p. 83).

Wang, W. et al. (July 2015). "Proteomic Analysis of Murine Testes Lipid Droplets". In: *Scientific reports* 5, p. 12070. DOI: 10.1038/srep12070. pmid: 26159641 (cit. on p. 83).

Bolcun-Filas, E. et al. (Aug. 2011). "A-MYB (MYBL1) Transcription Factor Is a Master Regulator of Male Meiosis". In: *Development* 138.15, pp. 3319–3330. DOI: 10.1242/dev.067645. pmid: 21750041 (cit. on pp. 83, 94).

Kojima, M. L., D. G. de Rooij, and D. C. Page (Feb. 2019). "Amplification of a Broad Transcriptional Program by a Common Factor Triggers the Meiotic Cell Cycle in Mice". In: *eLife* 8. DOI: 10.7554/eLife.43738. pmid: 30810530 (cit. on pp. 83, 85, 94).

Kistler, W. S. et al. (July 2015). "RFX2 Is a Major Transcriptional Regulator of Spermiogenesis". In: *PLoS genetics* 11.7, e1005368. DOI: 10.1371/journal.pgen.1005368. pmid: 26162102 (cit. on pp. 83, 94).

Nantel, F. et al. (Mar. 1996). "Spermiogenesis Deficiency and Germ-Cell Apoptosis in CREM-Mutant Mice". In: *Nature* 380.6570, pp. 159–162. DOI: 10.1038/380159a0 (cit. on p. 83).

Rosa, A. et al. (Dec. 2007). "The Interplay between the Master Transcription Factor PU.1 and miR-424 Regulates Human Monocyte/Macrophage Differentiation". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.50, pp. 19849–19854. DOI: 10.1073/pnas.0706963104. pmid: 18056638 (cit. on p. 85).

Li, E. and Y. Zhang (Jan. 5, 2014). "DNA Methylation in Mammals". In: *Cold Spring Harbor Perspectives in Biology* 6.5, a019133. DOI: 10.1101/cshperspect.a019133. pmid: 24789823 (cit. on p. 87).

Domcke, S. et al. (Dec. 2015). "Competition between DNA Methylation and Transcription Factors Determines Binding of NRF1". In: *Nature* 528.7583, pp. 575–579. DOI: 10.1038/nature16462. pmid: 26675734 (cit. on p. 87).

Wang, J. et al. (Nov. 2017). "NRF1 Coordinates with DNA Methylation to Regulate Spermatogenesis". In: *The FASEB Journal* 31.11, pp. 4959–4970. DOI: 10.1096/fj.201700093r (cit. on p. 87).

Imai, Y. et al. (Dec. 2017). "The PRDM9 KRAB Domain Is Required for Meiosis and Involved in Protein Interactions". In: *Chromosoma* 126.6, pp. 681–695. DOI: 10.1007/s00412-017-0631-z. pmid: 28527011 (cit. on p. 87).

Parvanov, E. D. et al. (Feb. 2017). "PRDM9 Interactions with Other Proteins Provide a Link between Recombination Hotspots and the Chromosomal Axis in Meiosis". In: *Molecular biology of the cell* 28.3, pp. 488–499. DOI: 10.1091/mbc.E16-09-0686. pmid: 27932493 (cit. on pp. 87, 119).

Banito, A. et al. (Mar. 2018). "The SS18-SSX Oncoprotein Hijacks KDM2B-PRC1.1 to Drive Synovial Sarcoma". In: *Cancer cell* 33.3, 527–541.e8. DOI: 10.1016/j.ccell.2018.01.018. pmid: 29502955 (cit. on pp. 87, 116).

Long, H. K., N. P. Blackledge, and R. J. Klose (June 1, 2013). "ZF-CxxC Domain-Containing Proteins, CpG Islands and the Chromatin Connection". In: *Biochemical Society Transactions* 41.3, pp. 727–740. DOI: 10.1042/BST20130028 (cit. on p. 87).

Lee, J.-H. and D. G. Skalnik (Dec. 16, 2005). "CpG-Binding Protein (CXXC Finger Protein 1) Is a Component of the Mammalian Set1 Histone H3-Lys4 Methyltransferase Complex, the Analogue of the Yeast Set1/COMPASS Complex". In: *Journal of Biological Chemistry* 280.50, pp. 41725–41731. DOI: 10.1074/jbc.M508312200. pmid: 16253997 (cit. on p. 87).

He, J. et al. (Nov. 2008). "The H3K36 Demethylase Jhdm1b/Kdm2b Regulates Cell Proliferation and Senescence through P15 Ink4b". In: *Nature Structural & Molecular Biology* 15.11, pp. 1169–1175. DOI: 10.1038/nsmb.1499 (cit. on p. 87).

Farcas, A. M. et al. (Dec. 2012). "KDM2B Links the Polycomb Repressive Complex 1 (PRC1) to Recognition of CpG Islands". In: *eLife* 1. DOI: 10.7554/elife.00205 (cit. on p. 87).

He, J. et al. (Apr. 2013). "Kdm2b Maintains Murine Embryonic Stem Cell Status by Recruiting PRC1 Complex to CpG Islands of Developmental Genes". In: *Nature Cell Biology* 15.4, pp. 373–384. DOI: 10.1038/ncb2702 (cit. on p. 87).

Wu, X., J. V. Johansen, and K. Helin (Mar. 28, 2013). "Fbxl10/Kdm2b Recruits Polycomb Repressive Complex 1 to CpG Islands and Regulates H2A Ubiquitylation". In: *Molecular Cell* 49.6, pp. 1134–1146. DOI: 10.1016/j.molcel.2013.01.016. pmid: 23395003 (cit. on p. 87).

Sin, H.-S. et al. (July 2015). "Poised Chromatin and Bivalent Domains Facilitate the Mitosis-to-Meiosis Transition in the Male Germline". In: *BMC biology* 13, p. 53. DOI: 10.1186/s12915-015-0159-8. pmid: 26198001 (cit. on p. 87).

Sassone-Corsi, P. (2000). "CREM: A Master-Switch Regulating the Balance between Differentiation and Apoptosis in Male Germ Cells". In: *Molecular reproduction and development* 56.S2, pp. 228–229. DOI: 10.1002/(sici)1098-2795(200006)56:2+<228::aid-mrd2>3.0.co;2-b (cit. on p. 88).

Dobin, A. et al. (Jan. 2013). "STAR: Ultrafast Universal RNA-Seq Aligner". In: *Bioinformatics* 29.1, pp. 15–21. DOI: 10.1093/bioinformatics/bts635. pmid: 23104886 (cit. on p. 90).

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing (cit. on pp. 90, 131).

Yu, G. et al. (May 2012). "clusterProfiler: An R Package for Comparing Biological Themes among Gene Clusters". In: *Omics: a journal of integrative biology* 16.5, pp. 284–287. DOI: 10.1089/omi.2011.0118. pmid: 22455463 (cit. on p. 90).
Campitelli, E. (2019). *Ggnewscale: Multiple Fill and Color Scales in 'ggplot2'*. DOI: 10.5281/zenodo.2543762 (cit. on p. 90).
Garnier, S. (2018). *Viridis: Default Color Maps from 'Matplotlib'* (cit. on pp. 90, 132).
Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes* (cit. on pp. 90, 132).
Pedersen, T. L. (2016). *Ggforce: Accelerating 'Ggplot2'* (cit. on pp. 90, 132).
Petukhov, V. (2018). *Ggrastr: Raster Layers for Ggplot2* (cit. on p. 90).
Wagih, O. (Nov. 2017). "Ggseqlogo: A Versatile R Package for Drawing Sequence Logos". In: *Bioinformatics* 33.22, pp. 3645–3647. DOI: 10.1093/bioinformatics/btx469. pmid: 29036507 (cit. on p. 90).
Wickham, H. (June 2016). *Ggplot2: Elegant Graphics for Data Analysis*. Springer (cit. on pp. 90, 132).
Wilke, C. O. (2018). *Cowplot: Streamlined Plot Theme and Plot Annotations for 'Ggplot2'* (cit. on pp. 90, 132).
Bates, D. and M. Maechler (2018). *Matrix: Sparse and Dense Matrix Classes and Methods* (cit. on p. 91).
Chang, W. et al. (2018). *Shiny: Web Application Framework for R* (cit. on p. 91).
Dowle, M. and A. Srinivasan (2019). *Data.Table: Extension of 'data.Frame'* (cit. on pp. 91, 132).
Durinck, S. et al. (Aug. 2005). "BioMart and Bioconductor: A Powerful Link between Biological Databases and Microarray Data Analysis". In: *Bioinformatics* 21.16, pp. 3439–3440. DOI: 10.1093/bioinformatics/bti525. pmid: 16082012 (cit. on p. 91).
Gu, Z., R. Eils, and M. Schlesner (Sept. 2016). "Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data". In: *Bioinformatics* 32.18, pp. 2847–2849. DOI: 10.1093/bioinformatics/btw313. pmid: 27207943 (cit. on pp. 91, 132).
Kane, M. J., J. Emerson, and S. Weston (2013). "Scalable Strategies for Computing with Massive Data". In: 55. DOI: 10.18637/jss.v055.i14 (cit. on p. 91).
Sali, A. (2017). *Shinycssloaders: Add CSS Loading Animations to 'shiny' Outputs* (cit. on p. 91).
Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. New York: Springer (cit. on p. 91).
Lin, X. and P. C. Boutros (2019). *NNLM: Fast and Versatile Non-Negative Matrix Factorization* (cit. on pp. 91, 92).
Oksanen, J. et al. (2019). *Vegan: Community Ecology Package* (cit. on p. 91).
Marchini, J. L., C. Heaton, and B. D. Ripley (2017). *fastICA: FastICA Algorithms to Perform ICA and Projection Pursuit* (cit. on p. 92).
Abraham, G. and M. Inouye (Apr. 2014). "Fast Principal Component Analysis of Large-Scale Genome-Wide Data". In: *PloS one* 9.4, e93766. DOI: 10.1371/journal.pone.0093766. pmid: 24718290 (cit. on p. 92).
Bailey, T. L. et al. (July 2015). "The MEME Suite". In: *Nucleic acids research* 43.W1, W39–49. DOI: 10.1093/nar/gkv416. pmid: 25953851 (cit. on p. 93).
Gupta, S. et al. (Feb. 2007). "Quantifying Similarity between Motifs". In: *Genome Biology* 8.2, R24+. DOI: 10.1186/gb-2007-8-2-r24. pmid: 17324271 (cit. on p. 93).

Kulakovskiy, I. V. et al. (Jan. 2018). "HOCOMOCO: Towards a Complete Collection of Transcription Factor Binding Models for Human and Mouse via Large-Scale ChIP-Seq Analysis". In: *Nucleic acids research* 46.D1, pp. D252–D259. DOI: 10.1093/nar/gkx1106. pmid: 29140464 (cit. on p. 93).

GTEx Consortium (June 2013). "The Genotype-Tissue Expression (GTEx) Project". In: *Nature genetics* 45.6, pp. 580–585. DOI: 10.1038/ng.2653. pmid: 23715323 (cit. on p. 93).

Kosir, R. et al. (Feb. 2012). "Novel Insights into the Downstream Pathways and Targets Controlled by Transcription Factors CREM in the Testis". In: *PloS one* 7.2, e31798. DOI: 10.1371/journal.pone.0031798 (cit. on p. 94).

Sutton, W. S. (1902). "On the Morphology of the Chromoso Group in Brachystola Magna". In: *The Biological Bulletin* 4.1, pp. 24–39 (cit. on p. 95).

He, F., Y. Muto, et al. (2010). "Complex Structure of the Zf-CW Domain and the H3K4me3 Peptide". In: DOI: 10.2210/pdb2rr4/pdb (cit. on p. 96).

Rona, G. B., E. C. A. Eleutherio, and A. S. Pinheiro (Mar. 2016). "PWWP Domains and Their Modes of Sensing DNA and Histone Methylated Lysines". In: *Biophysical reviews* 8.1, pp. 63–74. DOI: 10.1007/s12551-015-0190-6. pmid: 28510146 (cit. on p. 96).

Carithers, L. J. et al. (Oct. 2015). "A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project". In: *Biopreservation and biobanking* 13.5, pp. 311–319. DOI: 10.1089/bio.2015.0032. pmid: 26484571 (cit. on p. 96).

Li, M. et al. (2019). "The Histone Modification Reader ZCWPW1 Is Required for Meiosis Prophase I in Male but Not in Female Mice". In: *Science Advances* 5.8, eaax1101. DOI: 10.1126/sciadv.aax1101 (cit. on pp. 96, 100, 101, 104, 116, 117, 140, 141).

Wells, D. et al. (Aug. 3, 2020). "ZCWPW1 Is Recruited to Recombination Hotspots by PRDM9 and Is Essential for Meiotic Double Strand Break Repair". In: *eLife* 9. Ed. by B. de Massy, J. K. Tyler, and B. de Massy, e53392. DOI: 10.7554/eLife.53392 (cit. on p. 96).

Baker, Z. et al. (June 6, 2017). "Repeated Losses of PRDM9-Directed Recombination despite the Conservation of PRDM9 across Vertebrates". In: *eLife* 6. Ed. by B. de Massy, e24133. DOI: 10.7554/eLife.24133 (cit. on pp. 97–99, 121, 122).

Lobley, A., M. I. Sadowski, and D. T. Jones (July 2009). "pGenTHREADER and pDomTHREADER: New Methods for Improved Protein Fold Recognition and Superfamily Discrimination". In: *Bioinformatics* 25.14, pp. 1761–1767. DOI: 10.1093/bioinformatics/btp302. pmid: 19429599 (cit. on pp. 97, 98, 122).

Marchler-Bauer, A. and S. H. Bryant (July 2004). "CD-Search: Protein Domain Annotations on the Fly". In: *Nucleic acids research* 32 (Web Server issue), W327–31. DOI: 10.1093/nar/gkh454. pmid: 15215404 (cit. on p. 98).

Capra, J. A. and M. Singh (Aug. 1, 2007). "Predicting Functionally Important Residues from Sequence Conservation". In: *Bioinformatics* 23.15, pp. 1875–1882. DOI: 10.1093/bioinformatics/btm270 (cit. on p. 98).

Johansson, F. and H. Toh (July 21, 2010). "A Comparative Study of Conservation and Variation Scores". In: *BMC Bioinformatics* 11.1, p. 388. DOI: 10.1186/1471-2105-11-388 (cit. on p. 98).

Thibault-Sennett, S. et al. (June 2018). "Interrogating the Functions of PRDM9 Domains in Meiosis". In: *Genetics* 209.2, pp. 475–487. DOI: 10.1534/genetics.118.300565. pmid: 29674518 (cit. on p. 99).

Tarsounas, M. et al. (Oct. 1999). "RAD51 and DMC1 Form Mixed Complexes Associated with Mouse Meiotic Chromosome Cores and Synaptonemal Complexes". In: *The Journal of cell biology* 147.2, pp. 207–220. DOI: 10.1083/jcb.147.2.207. pmid: 10525529 (cit. on p. 104).

Petukhova, G. V., P. J. Romanienko, and R. D. Camerini-Otero (Dec. 1, 2003). "The Hop2 Protein Has a Direct Role in Promoting Interhomolog Interactions during Mouse Meiosis". In: *Developmental Cell* 5.6, pp. 927–936. DOI: 10.1016/S1534-5807(03)00369-1 (cit. on p. 113).

Deininger, P. (Dec. 28, 2011). "Alu Elements: Know the SINEs". In: *Genome Biology* 12.12, p. 236. DOI: 10.1186/gb-2011-12-12-236 (cit. on p. 113).

Luo, Y., X. Lu, and H. Xie (2014). "Dynamic Alu Methylation during Normal Development, Aging, and Tumorigenesis". In: *BioMed Research International*. DOI: 10.1155/2014/784706. pmid: 25243180 (cit. on p. 114).

Cahoon, C. K. and D. E. Libuda (Mar. 2019). "Leagues of Their Own: Sexually Dimorphic Features of Meiotic Prophase I". In: *Chromosoma*. DOI: 10.1007/s00412-019-00692-x. pmid: 30826870 (cit. on p. 116).

Morelli, M. A. and P. E. Cohen (Dec. 2005). "Not All Germ Cells Are Created Equal: Aspects of Sexual Dimorphism in Mammalian Meiosis". In: *Reproduction* 130.6, pp. 761–781. DOI: 10.1530/rep.1.00865. pmid: 16322537 (cit. on p. 116).

Ashley, T., D. Walpita, and D. G. de Rooij (Feb. 2001). "Localization of Two Mammalian Cyclin Dependent Kinases during Mammalian Meiosis". In: *Journal of cell science* 114 (Pt 4), pp. 685–693. pmid: 11171374 (cit. on p. 117).

Horn, H. F. et al. (Sept. 2013). "A Mammalian KASH Domain Protein Coupling Meiotic Chromosomes to the Cytoskeleton". In: *The Journal of cell biology* 202.7, pp. 1023–1039. DOI: 10.1083/jcb.201304004. pmid: 24062341 (cit. on p. 117).

Shibuya, H., K.-I. Ishiguro, and Y. Watanabe (Feb. 2014). "The TRF1-Binding Protein TERB1 Promotes Chromosome Movement and Telomere Rigidity in Meiosis". In: *Nature cell biology* 16.2, pp. 145–156. DOI: 10.1038/ncb2896. pmid: 24413433 (cit. on p. 117).

Schultz, D. C. et al. (Apr. 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". In: *Genes & development* 16.8, pp. 919–932. DOI: 10.1101/gad.973302. pmid: 11959841 (cit. on p. 119).

Mulligan, P. et al. (Dec. 2008). "CDYL Bridges REST and Histone Methyltransferases for Gene Repression and Suppression of Cellular Transformation". In: *Molecular cell* 32.5, pp. 718–726. DOI: 10.1016/j.molcel.2008.10.025. pmid: 19061646 (cit. on p. 119).

Seisenberger Stefanie et al. (Jan. 5, 2013). "Reprogramming DNA Methylation in the Mammalian Life Cycle: Building and Breaking Epigenetic Barriers". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1609, p. 20110330. DOI: 10.1098/rstb.2011.0330 (cit. on p. 119).

Liu, Y., W. Tempel, et al. (Apr. 22, 2016). "Family-Wide Characterization of Histone Binding Abilities of Human CW Domain-Containing Proteins". In: *Journal of Biological Chemistry* 291.17, pp. 9000–9013. DOI: 10.1074/jbc.M116.718973. pmid: 26933034 (cit. on p. 119).

Altschul, S. F. et al. (Sept. 1997). "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs". In: *Nucleic acids research* 25.17, pp. 3389–3402. DOI: 10.1093/nar/25.17.3389. pmid: 9254694 (cit. on p. 120).

Altschul, S. F. et al. (Oct. 2005). "Protein Database Searches Using Compositionally Adjusted Substitution Matrices". In: *The FEBS journal* 272.20, pp. 5101–5109. DOI: 10.1111/j.1742-4658.2005.04945.x. pmid: 16218944 (cit. on p. 120).

He, F., T. Umehara, et al. (Sept. 8, 2010). "Structural Insight into the Zinc Finger CW Domain as a Histone Modification Reader". In: *Structure* 18.9, pp. 1127–1139. DOI: 10.1016/j.str.2010.06.012. pmid: 20826339 (cit. on p. 121).

Qin, S. and J. Min (Nov. 2014). "Structure and Function of the Nucleosome-Binding PWWP Domain". In: *Trends in biochemical sciences* 39.11, pp. 536–547. DOI: 10.1016/j.tibs.2014.09.001. pmid: 25277115 (cit. on p. 121).

Barchi, M. et al. (May 2008). "ATM Promotes the Obligate XY Crossover and Both Crossover Control and Chromosome Axis Integrity on Autosomes". In: *PLoS genetics* 4.5, e1000076. DOI: 10.1371/journal.pgen.1000076. pmid: 18497861 (cit. on p. 124).

Peters, A. H. et al. (Feb. 1997). "A Drying-down Technique for the Spreading of Mammalian Meiocytes from the Male and Female Germline". In: *Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* 5.1, pp. 66–68. pmid: 9088645 (cit. on p. 124).

Campeau, E. et al. (Aug. 6, 2009). "A Versatile Viral System for Expression and Depletion of Proteins in Mammalian Cells". In: *PLOS ONE* 4.8, e6529. DOI: 10.1371/journal.pone.0006529 (cit. on p. 125).

Lunter, G. and M. Goodson (June 2011). "Stampy: A Statistical Algorithm for Sensitive and Fast Mapping of Illumina Sequence Reads". In: *Genome research* 21.6, pp. 936–939. DOI: 10.1101/gr.111120.110. pmid: 20980556 (cit. on p. 129).

Libertini, E. et al. (2015). "Overexpression of the Heterochromatinization Factor BAHD1 in HEK293 Cells Differentially Reshapes the DNA Methylome on Autosomes and X Chromosome". In: *Frontiers in Genetics* 6. DOI: 10.3389/fgene.2015.00339 (cit. on pp. 131, 132).

Pratto, F. et al. (Nov. 2014b). "DNA Recombination. Recombination Initiation Maps of Individual Human Genomes". In: *Science* 346.6211, p. 1256442. DOI: 10.1126/science.1256442. pmid: 25395542 (cit. on pp. 131, 132).

Köster, J. and S. Rahmann (Oct. 1, 2012). "Snakemake—a Scalable Bioinformatics Workflow Engine". In: *Bioinformatics* 28.19, pp. 2520–2522. DOI: 10.1093/bioinformatics/bts480 (cit. on p. 131).

Pohl, A. and M. Beato (June 1, 2014). "Bwtool: A Tool for bigWig Files". In: *Bioinformatics* 30.11, pp. 1618–1619. DOI: 10.1093/bioinformatics/btu056 (cit. on p. 131).

Quinlan, A. R. and I. M. Hall (Mar. 15, 2010). "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features". In: *Bioinformatics* 26.6, pp. 841–842. DOI: 10.1093/bioinformatics/btq033 (cit. on p. 131).

Shen, W. et al. (Oct. 5, 2016). "SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation". In: *PLOS ONE* 11.10, e0163962. DOI: 10.1371/journal.pone.0163962 (cit. on p. 131).

Kent, W. J. et al. (Sept. 1, 2010). "BigWig and BigBed: Enabling Browsing of Large Distributed Datasets". In: *Bioinformatics* 26.17, pp. 2204–2207. DOI: 10.1093/bioinformatics/btq351 (cit. on p. 131).

Thorvaldsdóttir, H., J. T. Robinson, and J. P. Mesirov (Mar. 2013). "Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and

Exploration". In: *Briefings in Bioinformatics* 14.2, pp. 178–192. DOI: 10.1093/bib/bbs017. pmid: 22517427 (cit. on p. 131).

Tange, O. (Feb. 2011). "GNU Parallel: The Command-Line Power Tool". In: p. 6 (cit. on p. 131).

Brennan, P. (July 18, 2018). "drawProteins: A Bioconductor/R Package for Reproducible and Programmatic Generation of Protein Schematics". In: *F1000Research* 7, p. 1105. DOI: 10.12688/f1000research.14541.1 (cit. on p. 132).

Wood, S. N. (2011). "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 73.1, pp. 3–36 (cit. on p. 132).

Grau, J., I. Grosse, and J. Keilwagen (Aug. 1, 2015). "PRROC: Computing and Visualizing Precision-Recall and Receiver Operating Characteristic Curves in R". In: *Bioinformatics* 31.15, pp. 2595–2597. DOI: 10.1093/bioinformatics/btv153 (cit. on p. 132).

Hill, J. T. et al. (2014). "Poly Peak Parser: Method and Software for Identification of Unknown Indels Using Sanger Sequencing of Polymerase Chain Reaction Products". In: *Developmental Dynamics* 243.12, pp. 1632–1636. DOI: 10.1002/dvdy.24183 (cit. on p. 132).

Janssens, F. A. (1909). "La Théorie de La Chiasmatypie. Nouvelle Interprétation Des Cinèses de Maturation (The Chiasmatype Theory. A New Interpretation of the Maturation Divisions.)" In: *Cellule* 25, pp. 389–411 (cit. on p. 133).

Hermann, B. P. et al. (Nov. 2018). "The Mammalian Spermatogenesis Single-Cell Transcriptome, from Spermatogonial Stem Cells to Spermatids". In: *Cell reports* 25.6, 1650–1667.e8. DOI: 10.1016/j.celrep.2018.10.026. pmid: 30404016 (cit. on p. 134).

Ernst, C. et al. (Mar. 19, 2019). "Staged Developmental Mapping and X Chromosome Transcriptional Dynamics during Mouse Spermatogenesis". In: *Nature Communications* 10.1, pp. 1–20. DOI: 10.1038/s41467-019-09182-1 (cit. on pp. 134, 137–139).

Sohni, A. et al. (Feb. 5, 2019). "The Neonatal and Adult Human Testis Defined at the Single-Cell Level". In: *Cell Reports* 26.6, 1501–1517.e4. DOI: 10.1016/j.celrep.2019.01.045 (cit. on pp. 134, 138).

Grive, K. J. et al. (Mar. 20, 2019). "Dynamic Transcriptome Profiles within Spermatogonial and Spermatocyte Populations during Postnatal Testis Maturation Revealed by Single-Cell Sequencing". In: *PLOS Genetics* 15.3, e1007810. DOI: 10.1371/journal.pgen.1007810 (cit. on pp. 134, 137).

Law, N. C., M. J. Oatley, and J. M. Oatley (June 26, 2019). "Developmental Kinetics and Transcriptome Dynamics of Stem Cell Specification in the Spermatogenic Lineage". In: *Nature Communications* 10.1, pp. 1–14. DOI: 10.1038/s41467-019-10596-0 (cit. on pp. 134, 138).

La, H. M. et al. (July 19, 2018). "Identification of Dynamic Undifferentiated Cell States within the Male Germline". In: *Nature Communications* 9.1, pp. 1–18. DOI: 10.1038/s41467-018-04827-z (cit. on p. 134).

Fang, X. et al. (Oct. 15, 2019). "Proteomics and Single-Cell RNA Analysis of Akap4-Knockout Mice Model Confirm Indispensable Role of Akap4 in Spermatogenesis". In: *Developmental Biology* 454.2, pp. 118–127. DOI: 10.1016/j.ydbio.2019.06.017 (cit. on pp. 134, 138).

Guo, J. et al. (Dec. 2018). "The Adult Human Testis Transcriptional Cell Atlas". In: *Cell Research* 28.12, pp. 1141–1157. DOI: 10.1038/s41422-018-0099-2 (cit. on pp. 133, 134).

Xia, B. et al. (Jan. 22, 2019). "Widespread Transcriptional Scanning in the Testis Modulates Gene Evolution Rates". In: *bioRxiv*, p. 282129. DOI: 10.1101/282129 (cit. on pp. 134, 137, 138).

Wang, M. et al. (Oct. 4, 2018). "Single-Cell RNA Sequencing Analysis Reveals Sequential Cell Fate Transition during Human Spermatogenesis". In: *Cell Stem Cell* 23.4, 599–614.e4. DOI: 10.1016/j.stem.2018.08.007 (cit. on pp. 134, 138).

Lukassen, S. et al. (Apr. 2018). "Characterization of Germ Cell Differentiation in the Male Mouse through Single-Cell RNA Sequencing". In: *Scientific reports* 8.1, p. 6521. DOI: 10.1038/s41598-018-24725-0. pmid: 29695820 (cit. on pp. 134, 138).

Vértesy, Á. et al. (July 8, 2019). "Dynamics of Meiotic Sex Chromosome Inactivation and Pachytene Activation in Mice Spermatogenesis". In: *bioRxiv*, p. 665372. DOI: 10.1101/665372 (cit. on pp. 134, 135).

Chen, Y. et al. (Sept. 2018). "Single-Cell RNA-Seq Uncovers Dynamic Processes and Critical Regulators in Mouse Spermatogenesis". In: *Cell research* 28.9, pp. 879–896. DOI: 10.1038/s41422-018-0074-y. pmid: 30061742 (cit. on pp. 134, 135, 140).

Stévant, I. et al. (Feb. 6, 2018). "Deciphering Cell Lineage Specification during Male Sex Determination with Single-Cell RNA Sequencing". In: *Cell Reports* 22.6, pp. 1589–1599. DOI: 10.1016/j.celrep.2018.01.043. pmid: 29425512 (cit. on pp. 134, 135).

Song, H.-W. et al. (Sept. 27, 2016). "The Homeobox Transcription Factor RHOX10 Drives Mouse Spermatogonial Stem Cell Establishment". In: *Cell Reports* 17.1, pp. 149–164. DOI: 10.1016/j.celrep.2016.08.090. pmid: 27681428 (cit. on p. 134).

Makino, Y. et al. (Apr. 15, 2019). "Single Cell RNA-Sequencing Identified Dec2 as a Suppressive Factor for Spermatogonial Differentiation by Inhibiting Sohlh1 Expression". In: *Scientific Reports* 9.1, pp. 1–15. DOI: 10.1038/s41598-019-42578-z (cit. on p. 134).

Neuhaus, N. et al. (Feb. 10, 2017). "Single-Cell Gene Expression Analysis Reveals Diversity among Human Spermatogonia". In: *Molecular Human Reproduction* 23.2, pp. 79–90. DOI: 10.1093/molehr/gaw079 (cit. on p. 134).

Guo, J. et al. (Oct. 5, 2017). "Chromatin and Single-Cell RNA-Seq Profiling Reveal Dynamic Signaling and Metabolic Transitions during Human Spermatogonial Stem Cell Development". In: *Cell Stem Cell* 21.4, 533–546.e6. DOI: 10.1016/j.stem.2017.09.003 (cit. on pp. 133, 134).

Liao, J. et al. (Mar. 15, 2019). "Revealing Cellular and Molecular Transitions in Neonatal Germ Cell Differentiation Using Single Cell RNA Sequencing". In: *Development* 146.6. DOI: 10.1242/dev.174953. pmid: 30824552 (cit. on p. 134).

Brawley, C. and E. Matunis (May 28, 2004). "Regeneration of Male Germline Stem Cells by Spermatogonial Dedifferentiation in Vivo". In: *Science* 304.5675, pp. 1331–1334. DOI: 10.1126/science.1097676. pmid: 15143218 (cit. on p. 134).

Nakagawa, T. et al. (Apr. 2, 2010). "Functional Hierarchy and Reversibility Within the Murine Spermatogenic Stem Cell Compartment". In: *Science* 328.5974, pp. 62–67. DOI: 10.1126/science.1182868. pmid: 20299552 (cit. on p. 134).

Hara, K. et al. (May 1, 2014). "Mouse Spermatogenic Stem Cells Continually Interconvert between Equipotent Singly Isolated and Syncytial States". In: *Cell Stem Cell* 14.5, pp. 658–672. DOI: 10.1016/j.stem.2014.01.019. pmid: 24792118 (cit. on p. 134).

Oatley, J. M. et al. (Apr. 1, 2009). "Colony Stimulating Factor 1 Is an Extrinsic Stimulator of Mouse Spermatogonial Stem Cell Self-Renewal". In: *Development* 136.7, pp. 1191–1199. DOI: 10.1242/dev.032243. pmid: 19270176 (cit. on p. 134).

Yang, Q.-E. et al. (Feb. 15, 2013). "CXCL12–CXCR4 Signaling Is Required for the Maintenance of Mouse Spermatogonial Stem Cells". In: *Journal of Cell Science* 126.4, pp. 1009–1020. DOI: 10.1242/jcs.119826. pmid: 23239029 (cit. on p. 134).

Jantsch, V. et al. (Oct. 3, 2007). "Caenorhabditis Elegans Prom-1 Is Required for Meiotic Prophase Progression and Homologous Chromosome Pairing". In: *Molecular Biology of the Cell* 18.12, pp. 4911–4920. DOI: 10.1091/mbc.e07-03-0243 (cit. on p. 135).

Hua, R. et al. (Dec. 16, 2019). "FBXO47 Regulates Telomere-Inner Nuclear Envelope Integration by Stabilizing TRF2 during Meiosis". In: *Nucleic Acids Research* 47.22, pp. 11755–11770. DOI: 10.1093/nar/gkz992 (cit. on p. 135).

Feng, C.-W. A. et al. (Dec. 15, 2017). "SOX30 Is Required for Male Fertility in Mice". In: *Scientific Reports* 7.1, pp. 1–12. DOI: 10.1038/s41598-017-17854-5 (cit. on p. 135).

Hirota, T. et al. (Dec. 3, 2018). "SETDB1 Links the Meiotic DNA Damage Response to Sex Chromosome Silencing in Mice". In: *Developmental Cell* 47.5, 645–659.e6. DOI: 10.1016/j.devcel.2018.10.004. pmid: 30393076 (cit. on p. 137).

Schmidt, E. E. (July 1, 1996). "Transcriptional Promiscuity in Testes". In: *Current Biology* 6.7, pp. 768–769. DOI: 10.1016/S0960-9822(02)00589-4. pmid: 8805310 (cit. on p. 138).

Eddy, E. M., K. Toshimori, and D. A. O'Brien (2003). "Fibrous Sheath of Mammalian Spermatozoa". In: *Microscopy Research and Technique* 61.1, pp. 103–115. DOI: 10.1002/jemt.10320 (cit. on p. 138).

Chappell, L., A. J. Russell, and T. Voet (2018). "Single-Cell (Multi)Omics Technologies". In: *Annual Review of Genomics and Human Genetics* 19.1, pp. 15–41. DOI: 10.1146/annurev-genom-091416-035324. pmid: 29727584 (cit. on p. 139).

Hu, Y. et al. (2018). "Single Cell Multi-Omics Technology: Methodology and Application". In: *Frontiers in Cell and Developmental Biology* 6. DOI: 10.3389/fcell.2018.00028 (cit. on p. 139).

Stuart, T. and R. Satija (May 2019). "Integrative Single-Cell Analysis". In: *Nature Reviews Genetics* 20.5, pp. 257–272. DOI: 10.1038/s41576-019-0093-7 (cit. on p. 139).

Hériché, J.-K., S. Alexander, and J. Ellenberg (2019). "Integrating Imaging and Omics: Computational Methods and Challenges". In: *Annual Review of Biomedical Data Science* 2.1, pp. 175–197. DOI: 10.1146/annurev-biodatasci-080917-013328 (cit. on p. 139).

Argelaguet, R. et al. (Dec. 2019). "Multi-Omics Profiling of Mouse Gastrulation at Single-Cell Resolution". In: *Nature* 576.7787, pp. 487–491. DOI: 10.1038/s41586-019-1825-8 (cit. on p. 139).

Baron, C. S. and A. van Oudenaarden (Dec. 2019). "Unravelling Cellular Relationships during Development and Regeneration Using Genetic Lineage Tracing". In: *Nature Reviews Molecular Cell Biology* 20.12, pp. 753–765. DOI: 10.1038/s41580-019-0186-3 (cit. on p. 140).

McKenna, A. and J. A. Gagnon (June 15, 2019). "Recording Development with Single Cell Dynamic Lineage Tracing". In: *Development* 146.12. DOI: 10.1242/dev.169730. pmid: 31249005 (cit. on p. 140).

Mahgoub, M. et al. (Oct. 29, 2019). "Dual Histone Methyl Reader ZCWPW1 Facilitates Repair of Meiotic Double Strand Breaks". In: *bioRxiv*, p. 821603. DOI: 10.1101/821603 (cit. on p. 140).

Huang, T. et al. (Nov. 8, 2019). "The Histone Modification Reader ZCWPW1 Links Histone Methylation to Repair of PRDM9-Induced Meiotic Double Stand Breaks". In: *bioRxiv*, p. 836023. DOI: 10.1101/836023 (cit. on p. 141).

Muller, H. J. (1932). "Some Genetic Aspects of Sex". In: *The American Naturalist* 66.703, pp. 118–138 (cit. on p. 145).

Montgomery, T. H. (1901). "A Study of the Chromosomes of the Germ Cells of Metazoa". In: *Transactions of the American Philosophical Society* 20.2, pp. 154–236. DOI: 10.2307/1005428 (cit. on p. 161).